INVESTIGACION Y CIENCIA

Edición en español de **SCIENTIFIC AMERICAN**



ECOLOGIA DE LOS ESCARABAJOS ESTERCOLEROS

Enero 1980

200 Ptas.

Número 40
Enero 1980

INVESTIGACION Y CIENCIA

Edición en español de **SCIENTIFIC AMERICAN**

EDITA: Prensa Científica, S.A.
Calabria, 235-239
Barcelona-29 (ESPAÑA)

Colaboradores de este número:

Asesoramiento y traducción:

Raimundo Goberna: *Causas de la diabetes;* Miquel Gich: *Aleaciones con memoria de la forma;* Luisa Vilageliu: *Teoría neutralista de la evolución molecular;* Manuel Puigcerver: *Las galaxias primitivas;* Joandomènes Ros: *Ecología de los escarabajos estercoleros africanos;* Pedro Pascual: *Teoría cuántica y realidad;* Antonio Blanco: *Una estación neolítica y de la Edad de Hierro en una colina inglesa;* Luis Bou: *Juegos matemáticos;* E. Myro y Alicia Menéndez: *Taller y laboratorio.*

Ciencia y sociedad:

Miguel Aguilar

Libros:

Antonio Alabau, Clemente Sánchez-Garnica y Mercedes Durfort



LA PORTADA

La ilustración de la portada se centra en una bola de estiércol de elefante que ha sido moldeada por un macho de escarabajo pelotero de la especie *Kheper platynotus,* que es una de las más de 2000 especies de escarabajos estercoleros activos en Africa oriental (véase "La ecología de los escarabajos estercoleros africanos", de Bernd Heinrich y George A. Bartholomew, en este número). El macho es el insecto de la parte inferior izquierda; ha empezado a hacer rodar la bola acabada a lo largo de una distancia de varios metros, hasta un lugar donde la enterrará. Empuja la pelota con sus patas traseras al tiempo que se mantiene sobre la cabeza. Encima de la pelota se encuentra una hembra de la especie; será transportada de forma pasiva y se hundirá en el suelo con la bola cuando el macho excave un agujero con este fin. Una vez enterrado el ovillo, se alimentan del mismo y se aparean. La hembra pone un huevo, y el resto sirve de alimento para la larva.

Suscripciones:
Prensa Científica, S.A.
Calabria 235-239
Barcelona-29 (España)
Teléfono 322 05 51 ext. 41

Condiciones de suscripción:
España:
Un año (12 números): 2200 pesetas
Extranjero:
Un año (12 números): 36 U.S.$
Ejemplar atrasado ordinario:
225 pesetas
Ejemplar atrasado extraordinario:
340 pesetas

Distribución para España:
Distribuciones de Enlace, S.A.
Ausias March, 49, Barcelona-10

Distribución para los restantes países:
Editorial Labor, S.A.
Calabria, 235-239, Barcelona-29

Publicidad:
Madrid:
Gustavo Martínez Ovin
Avda. de Moratalaz, 137, Madrid-30
Tel. 430 84 81
Cataluña:
Miguel Munill
Balmes, 191, 2.º, 2.ª, Barcelona-6
Tels. 218 44 45 y 218 40 86

Controlado por O.J.D.



PROCEDENCIA DE LAS ILUSTRACIONES

Diseño de la portada de Tom Prentiss

Página	Fuente	Página	Fuente
7-13	Manuel Blanco y Miguel Alonso	46-55	Adolph E. Brotman
17	A. Bennett Jenson y Kozaburo Hayashi, National Institute of Dental Research (*arriba*); Takashi Onodera, National Institute of Dental Research (*abajo*)	59-61	Gabor Kiss
18-20	Patricia J. Wynne	62	Gabor Kiss; Kitt Peak National Observatory (*abajo, derecha*)
21	Joseph R. Williamson, Facultad de Medicina de la Universidad de Washington	63-67	Gabor Kiss
22-26	Patricia J. Wynne	71-72	Bernd Heinrich, Universidad de California en Berkeley
27	A. Bennett Jenson	73-77	Tom Prentiss
32-33	Goodyear Aerospace Corporation	81-92	Jerome Kuhl
34-41	Dan Todd	96	P. W. Dixon, Universidad de Nottingham
		98-101	Alan D. Iselin
		102-103	P. W. Dixon
		107-111	Ilil Arbel
		115	Stuart Travis
		116-119	Michael Goodman

ISSN 0210-136X
Dep. legal: B. 38.999-76
Fotocomposición Tecfa
Guizpúzcoa, 36 (local 1) Barcelona-20
Cayfosa. Santa Perpetua de Moguda
Barcelona
Printed in Spain - Impreso en España

Los autores

MANUEL BLANCO ("Reparación del material genético") es jefe del Laboratorio de Genética Microbiana del Instituto de Investigaciones Citológicas de la Caja de Ahorros de Valencia. Doctor ingeniero agrónomo, trabajó durante cinco años (1968-73) en el Laboratoire d'Enzymologie del Centre National de la Recherche Scientifique en Gif-sur-Yvette, Francia. Durante ese tiempo fue becario de la Fondation Joliot-Curie de París y de la Comunidad Económica Europea (EURATOM). Posteriormente ha sido contratado por dicho organismo europeo y por la Universidad del Estado de Río de Janeiro como profesor en cursos de radiobiología molecular. Sus investigaciones se centran en el estudio de la interacción entre los distintos procesos celulares en los que interviene el ácido desoxirribonucleico.

ABNER LOUIS NOTKINS ("Causas de la diabetes") es el jefe del laboratorio de medicina oral en el National Institute of Dental Research. Se graduó en el Yale College y se doctoró en medicina por la Universidad de Nueva York. Fue interno y residente en medicina interna en el John's Hopkins Hospital. Para completar su formación médica ingresó en los National Institutes of Health como investigador adscrito al National Cancer Institute. Se trasladó al National Institute of Dental Research en 1973. El trabajo de investigación de Notkins se ha centrado en los procesos inmunológicos implicados en las infecciones virales persistentes y recurrentes y, más recientemente, en los factores que controlan la latencia del herpes simple y la forma en que los virus causan la diabetes.

L. McDONALD SCHETKY ("Aleaciones con memoria de la forma") es director técnico de metalurgia en la International Copper Research Association de Nueva York. Estudió en el Rensselaer Polytechnic Institute, donde obtuvo su licenciatura en ingeniería química, el "master" en ingeniería metalúrgica y el doctorado en metalurgia física. De 1953 a 1956 trabajó como investigador en el Instituto de Tecnología de Massachusetts, dirigiendo la sección de investigación de materiales en el laboratorio de instrumentación de dicho Instituto. En 1956 fundó la Alloyd Corporation, compañía dedicada a la fabricación e investigación, especializada sobre todo en el estudio de materiales.

MOTOO KIMURA ("Teoría neutralista de la evolución molecular") dirige el departamento de genética de poblaciones del Instituto Nacional de Genética del Japón. Antes de graduarse trabajó en botánica en la Universidad de Kyoto, y en 1949 se unió al equipo de investigadores del Instituto de Genética, que acababa de fundarse. En 1953 se trasladó a los Estados Unidos, doctorándose en la Universidad de Wisconsin, en 1956, en el laboratorio de James F. Crow. Después volvió al Japón para continuar sus investigaciones en el Instituto de Genética. En 1976 el Emperador le concedió la Orden de la Cultura, el máximo reconocimiento cultural del Japón.

DAVID L. MEIER y RASHID A. SUNYAEV ("Las galaxias primitivas") son ambos astrónomos, el primero norteamericano y el segundo ruso. Meier es asociado postdoctoral en astrofísica teórica en el Instituto de Tecnología de California. Obtuvo sus grados de bachelor y master en física en la Universidad de Missouri en La Rolla y sus grados de master y doctor en astronomía en la Universidad de Texas en Austin, el último de ellos en 1977. Meier completó recientemente un período de asociado postdoctoral de la OTAN en el Instituto de Astronomía de la Universidad de Cambridge. Sunyaev es director del grupo de Procesos Elementales en Astrofísica en el Instituto de Investigación Cósmica de Moscú. Se graduó en el Instituto de Física Técnica de Moscú en 1966, recibió su doctorado en 1968 y en 1973 le fue otorgado el título de doctor en ciencias, todos ellos en astrofísica. Sunyaev ha trabajado en cosmología, astrofísica de rayos X y física del plasma. Meier y Sunyaev no se conocen personalmente, pero fueron presentados por Beatriz M. Tinsley, de la Universidad de Yale, por correspondencia. Trabajaron primero con ella en una publicación científica y después juntos en el presente artículo, comunicando sólo por correo y teléfono.

BERND HEINRICH y GEORGE A. BARTHOLOMEW ("La ecología de los escarabajos estercoleros africanos") estudian el comportamiento, la fisiología y la ecología de los insectos. Heinrich es profesor de entomología en la Universidad de California en Berkeley. Estudió zoología en la Universidad de Maine y obtuvo su doctorado por la Universidad de California en Los Angeles, en 1970, trabajando en el laboratorio de Bartholomew. Su trabajo se ha centrado en la termorregulación y en la energética de los insectos, y en el papel de la fisiología y el comportamiento para ayudar a comprender los mecanismos de la ecología y de la evolución. Heinrich publicó recientemente un libro ttitulado *Bumblebee Economics* ("Economía del abejorro";Harvard University Press). Bartholomew es profesor de zoología en la Universidad de California en Los Angeles. Se graduó en Berkeley y obtuvo su doctorado por la Universidad de Harvard en 1947. Desde entonces ha estado en la UCLA.

BERNARD D'ESPAGNAT ("Teoría cuántica y realidad") enseña física en la Universidad de París y dirige el Laboratorio de Física Teórica y de Partículas Elementales de Orsay. Hijo de Georges d'Espagnat, un pintor postimpresionista, se educó en la École Polytechnique y en la Sorbonne, donde obtuvo su doctorado en ciencias físicas en 1950. Después trabajó un año como adjunto de investigación con Enrico Fermi en la Universidad de Chicago y en 1953 obtuvo una beca para el Instituto Niels Bohr de Copenhague. El año siguiente fue invitado por Félix Bloch, director a la sazón de la Organización Europea para Investigaciones Nucleares (CERN), para crear, en Ginebra, el núcleo de lo que más tarde sería la división teórica de la Organización. Allí investigó sobre la teoría de las partículas con extrañeza y contribuyó al método de aplicar teoría de grupo al estudio de estos objetos. Entró a formar parte de la facultad de la Sorbonne en 1969. D'Espagnat ha escrito tres libros sobre fundamentos teóricos de la mecánica cuántica. El último, que llevará por título *A la recherche du réel* (*En busca de lo real*) se publicará pronto en francés.

P. W. DIXON ("Una estación neolítica y de la Edad de Hierro en lo alto de una colina inglesa") es conferenciante (*lecturer*) de arqueología medieval en la Universidad de Nottingham y conservador honorario del Museo de la Universidad. Estudió clásicas en el New College de Oxford y se doctoró después en arqueología con una tesis sobre las casas fortificadas y la sociedad bajomedievales en el norte de Inglaterra y en Escocia. Ha dirigido las excavaciones descritas en este artículo desde que comenzaron en 1969, y ha participado también en excavaciones de palacios de los Tudor en Greenwich y Richmond y en varias casas medievales.

Hace...

José M.ª López Piñero

...*cuatrocientos años*

Falleció Francisco Arceo, una de las grandes figuras de la cirugía europea del siglo XVI. Nacido en la localidad extremeña de Fregenal de la Sierra en 1493, estudió medicina en la Universidad de Alcalá. Trabajó después varios años en los hospitales del Monasterio de Guadalupe, que eran entonces un prestigioso centro de perfeccionamiento clínico para médicos que ya habían obtenido su título, además de tener una escuela para cirujanos. Desde los años veinte hasta el final de su vida, Arceo ejerció la profesión en las poblaciones extremeñas de Llerena, Fuente de Cantos y Badajoz, alcanzando extraordinaria fama como cirujano. Seis años antes de su muerte, publicó *De recta curandorum vulnerum ratione* (1574), que fue impresa en Amberes por Christophe Plantin gracias a la intervención del célebre humanista Benito Arias Montano, al que le unía una estrecha amistad. El libro alcanzó una gran difusión en toda Europa, siendo reimpreso en latín y traducido al inglés, francés, alemán y holandés.

La base doctrinal de la obra de Arceo es el galenismo de orientación avicenista, en una línea semejante a la *Practica in arte chirurgica copiosa* (1514) de Giovanni da Vigo, que es el cirujano más citado. El contenido clínico y operatorio es, sin embargo, muy superior al del tratado italiano. Está redactado en un latín muy cuidado, y su autor critica a los que utilizaban la lengua vulgar como favorecedores del intrusismo por parte de prácticos quirúrgicos sin formación académica. A este respecto, la postura de Arceo es el polo opuesto de la de otras grandes figuras de la cirugía española de esta época, como Dionisio Daza Chacón, que escribieron sus libros en romance con la intención expresa de mejorar la preparación de los cirujanos propiamente dichos, que desconocían el latín.

El tratado de Arceo incluye un elevado número de historias clínicas que son brillantes muestras del estilo expositivo y la objetividad características de la "observatio" médica renacentista. Todas ellas corresponden a casos procedentes de su propia práctica profesional. Lo mismo sucede con los criterios operatorios que defiende y las intervenciones quirúrgicas que describe. Entre estas últimas figura una de las más importantes aportaciones europeas a la cirugía plástica con anterioridad a Gaspare Tagliacozzi. Se trata de una aparatosa herida que se extendía desde las cejas hasta las comisuras de los labios, en forma de un colgajo con pedículo inferior que incluía el esqueleto nasal y la parte anterior de los maxilares superiores. Cuando el paciente acudió a Arceo, "la nariz y la mandíbula estaban frías, lacias y casi muertas, de modo que costaba mucho trabajo introducir la aguja". El cirujano extremeño repuso el colgajo en su sitio, suturó las partes blandas, ligó entre sí las piezas dentarias e inmovilizó el conjunto con un ingenioso dispositivo de vendas, "quedando las partes tan bien unidas y el apósito tan bien aplicado, que después de la curación sólo se conocía la cicatriz".

Otra importante contribución de Arceo es su estudio acerca de las alteraciones congénitas del pie en los niños, generalmente considerado como uno de los textos "clásicos" de la ortopedia. Describe a base de casuística propia sus principales tipos y propone un método original para su reducción ortopédica, tras reducir manualmente la articulación podálica. Incluye asimismo una figura con el modelo de bota ortopédica por él ideado.

Notable es también su exposición de las heridas cefálicas, a las que dedica varios capítulos, comenzando por las superficiales, para las que preconiza una cura por primera intención. Estudia detenidamente, apoyándose en su experiencia clínica, el diagnóstico y el tratamiento de las diferentes formas de fracturas craneales y sus complicaciones, principalmente las lesiones meníngeas y vasculares. Desde el punto de vista terapéutico, es partidario de la utilización del trépano, aunque critica agriamente a los que lo usaban innecesariamente o sin la suficiente pericia técnica. A remediar esta última limitación va dirigido un capítulo consagrado a un detenido análisis del trépano y de su forma de manejo, sin duda el más valioso de toda su exposición sobre el tema. El instrumento al que se refiere es el llamado "modiolus nespulatus", consistente en un vástago con sendas coronas de trépano en cada uno de sus extremos, que funcionaba haciéndolo girar entre las palmas de las manos. Las coronas llevaban topes adecuados para hacerlas insumergibles y tenían dos variantes: una con una lengüeta puntiaguda en su centro para iniciar la penetración ("trépano macho") y otra sin nada en el centro para concluirla ("trépano hembra").

Expone ampliamente la clínica y el tratamiento del cáncer de mama. Recomienda la extirpación radical de los "ocultos" con una técnica operatoria propia, y juzga que en los ya "ulcerados" sólo cabe una terapéutica puramente paliativa.

Menos interés tiene el resto de temas que examina en su obra, entre los que se encuentran las heridas torácicas, abdominales y de los miembros, las úlceras, la sífilis y las fiebres. A estas últimas dedica un estudio independiente titulado *De febrium curandorum ratione*.

...*cien años*

Se creó el Laboratorio Químico Municipal de Valencia –uno de los primeros de su clase en España– por iniciativa de José Monserrat y Riutort. Nacido en Valencia en 1814, Monserrat estudió en primer lugar medicina en su ciudad natal, acabando su carrera en la misma Valencia en 1835, y amplió después su formación en París. No obstante, sus inclinaciones le llevaron al campo de las ciencias físico-químicas, de las que alcanzó el doctorado en 1847. Con anterioridad había sido ayudante de la Cátedra de Química, cuyo titular le presentó al célebre científico francés Dumas, con el que mantuvo una activa relación personal. El mismo año 1847 consiguió la Cátedra de Química general, que desempeñó hasta su muerte. El prestigio de Monserrat dentro de la Universidad fue considerable, en especial en las Facultades de Ciencias y de Medicina, alcanzando sucesivamente el decanato de la primera, y el vicerrectorado y rectorado de la Universidad.

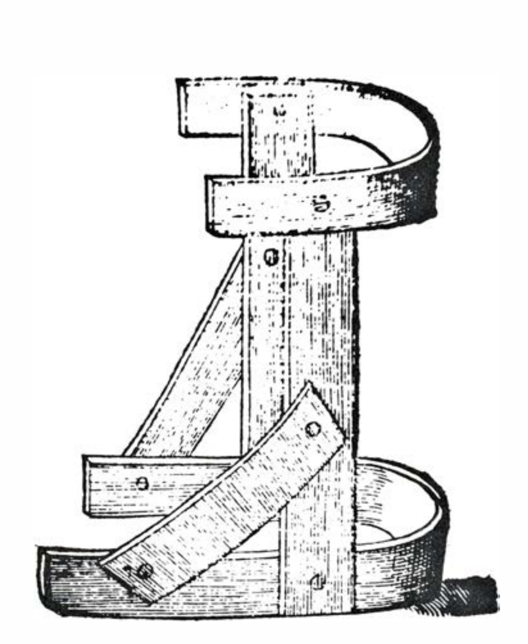
El interés de la labor de Monserrat reside en el hecho de que estuvo en el polo opuesto a las características habituales de la obra de los hombres más conocidos –y supuestamente más importantes– de la ciencia española de su tiempo. Les sobraba a estos últimos dotes de improvisación, erudición y tendencia a las generalizaciones más ambiciosas. Monserrat, por el contrario, fue un hombre de pretensiones modestas, pero que gastó su vida en el laboratorio. Es inevitable su comparación con una figura como Pedro Mata, autor de tratados de química casi sin apenas haber manejado en su vida un tubo de ensayo. El esfuerzo de Monserrat se centró en adquirir técnicas y hábitos de trabajos en enseñarlos a sus discípulos, en crear

las necesarias instituciones, y en que su ambiente –desde la medicina hasta la industria– se beneficiara de los mismos.

Su labor docente no se limitó a su cátedra y laboratorio, sino que dio cursos de análisis químico aplicado a las ciencias médicas y explicó como interino en la Escuela Industrial. Creó, durante su rectorado, enseñanzas prácticas especiales, entre las que destacaremos uno de los primeros cursos de investigación microscópica que se dieron en España. Mejoró también las instalaciones del Jardín Botánico y el Museo de Ciencias Naturales.

Monserrat fue el primero, junto a J. Gil, que aplicó la fotografía en España para fines científicos; en 1840, con la simple lectura de la memoria que acababa de publicar Daguerre en París, reconstruyó enteramente su aparato y su método. Más tarde lo aplicó, por ejemplo, a la obtención de fotografías astronómicas que llegaron a interesar a científicos extranjeros. Por otra parte, sintetizó durante mucho tiempo el cloroformo utilizado por los cirujanos valencianos y, asimismo, el hipoclorito y el ácido fénico para fines sanitarios. Introdujo también en España la fabricación de la sosa por el procedimiento Leblanc, la purificación moderna del arsénico, etc. El análisis químico se convirtió en sus manos en un instrumento para muy diversas necesidades: desde las aguas minero-medicinales y los productos farmacológicos, hasta los colorantes de los vinos y el dorado de la industria cerámica valenciana.

El número de sus publicaciones fue, en cambio, relativamente escaso. De todas ellas destaca su discurso inaugural de la Academia de Medicina de Valencia el año 1862, que constituye una pieza de importancia en la historia de la química fisiológica española. Se ocupa en ella del metabolismo de las proteínas, y además expone una teoría sobre la termogénesis y sobre el depósito de grasa. Bajo el influjo de Monserrat se cultivaron en Valencia tempranamente la bioquímica y la microbiología. De todos sus seguidores y discípulos destacan dos: Pablo Colvée Roura y Vicente Peset Cervera.







Modelo de bota ortopédica ideada por Francisco Arceo y portada de la edición de su obra en Amsterdam (1658), una de las ocho que a lo largo de un siglo tuvo en cinco idiomas.

Reparación del material genético

Reflejando la integración existente entre los procesos básicos de la célula, los mecanismos que reparan los genes aparecen estrechamente relacionados con los procesos que originan la mutagénesis y la activación de provirus

Manuel Blanco

El material genético de todos los organismos vivos, de la bacteria al hombre, está formado por una sustancia llamada ácido desoxirribonucleico, abreviadamente ADN. Este material genético determina, en gran parte, las características del organismo. Se transmite a los descendientes y es el responsable de que éstos presenten un parecido, más o menos grande, con sus progenitores. Para que esa transmisión sea posible, hace falta que las moléculas de ADN se dupliquen y den origen a nuevas moléculas, idénticas a la parental, que son las que heredarán las células hijas. El ADN encierra, pues, una información y tiene la capacidad de autoduplicarse.

Cuando Watson y Crick descubrieron, en 1953, la estructura del ADN, fue relativamente sencillo imaginar cómo esa molécula podía poseer propiedades tan extraordinarias. En efecto, observando la estructura del ADN encontramos en ella un verdadero mensaje escrito con un alfabeto de cuatro letras: A, G, C y T. Estas letras son las iniciales de los compuestos químicos, o bases nitrogenadas, adenina y guanina (bases purínicas), citosina y timina (bases pirimidínicas). Las bases forman parte de otros compuestos más complejos, los nucleótidos, que son los eslabones de las largas cadenas que constituyen el ADN. Las bases se disponen en secuencias (por ejemplo ATTCGAT... TGC), que determinan la estructura de las proteínas. Estas, a su vez, determinan las características del organismo (desde la fermentación del azúcar lactosa hasta tener los ojos de color marrón, por citar dos casos). Hay secuencias que no codifican ninguna proteína pero que tienen una función reguladora (verbigracia, influyen en la frecuencia con que se va a realizar la lectura del mensaje encerrado en otras secuencias).

Las cuatro bases citadas, además de constituir el mensaje genético, muestran la notable propiedad de acoplarse de un modo específico dos a dos. La adenina se acopla con la timina, formando el par A – T, y la guanina con la citosina, formando el par G – C. Esta propiedad de las bases hace que dos cadenas de nucleótidos puedan aparearse de un modo estable, adoptando, entonces, la forma de una escalera de caracol y constituyendo la llamada doble hélice. La estructura en forma de dos hélices, complementaria una de la otra, permite imaginar el mecanismo utilizado para la autoduplicación, es decir, para obtener dos moléculas hijas idénticas a la parental. Cada hélice de dicha molécula parental sirve de molde sobre el que se va construyendo la hélice complementaria. Se obtienen, así, dos moléculas hijas que son réplicas de la parental. Razón por la que, a este proceso, se le llama replicación del ADN.

Tal vez debido a que, a la vista de la estructura del ADN, resulta fácil comprender sus aspectos funcionales, es por lo que este componente celular básico ha sido considerado en ocasiones como algo inerte, que se limitaría a ser el programa fijo determinante de las características del organismo. En ese enfoque se apoyan quienes se resisten a creer que lo que podríamos llamar el puesto de mando de la célula sea algo que aparece como muerto. Además, es difícil conciliar la apariencia fija e invariable del ADN con el hecho de que sea, precisamente él, el material sobre el que trabajaría la evolución.

Que esta visión estática del ADN es muy limitada se ha hecho cada vez más evidente, en los últimos años, al conocerse su implicación en gran número de procesos celulares, de modo que puede hablarse de la existencia de un metabolismo del ADN, estrechamente relacionado con el metabolismo celular. Estos procesos del metabolismo del ADN –tales como la replicación, la transcripción, la recombinación, la reparación y otros– están sometidos a una regulación muy precisa y en ellos intervienen un grupo muy nutrido de enzimas y otras proteínas. Por codificarlas el propio ADN, resulta que una parte importante del mensaje genético de una célula determina procesos en los que va a estar implicado el propio soporte material de dicho mensaje.

El ADN se nos aparece como algo activo, capaz de interaccionar constantemente con su entorno físico-químico, y con posibilidades de aprovechar las ventajas y de sufrir los perjuicios que surjan de dicha interacción. Hablando hipotéticamente, podríamos decir que el ADN no se parece a un libro donde estarían escritas las "instrucciones" para la vida de la célula, pero que, en sí, sería algo muerto; podríamos decir, en cambio, que el ADN se parece a la memoria, la cual es el depósito de una información, pero un depósito flexible y vivo, un depósito –y esto es tal vez su característica más importante– creativo.

En este artículo vamos a ocuparnos de los procesos implicados en la reparación de las lesiones que se producen en el ADN. Las lesiones tienen un doble origen: espontáneo, por rotura al azar, o provocado, por someter a la célula a un tratamiento con cualquier agente físico o químico ante los que el ADN presenta una gran fragilidad. Esta fragilidad resulta de la estructura química de sus componentes, en especial de las bases nitrogenadas, que son muy reactivas y pueden ser modificadas por diversos tratamientos. La energía de la luz ultravioleta de longitud de onda de 260 nanómetros la absorben las bases pirimidínicas (citosina y timina). Cuando se da la circunstancia de que dos pirimidinas se hallan situadas contiguamente en la secuencia del ADN, utilizan dicha energía para formar enlaces covalentes (enlaces químicos fuertes) entre sí, resultando de esa unión los llamados dímeros de pirimidina. Las radiaciones de más energía (rayos X y rayos gamma) ionizan los átomos y, los iones formados, reaccio-

nan a su vez con las bases, alterando su estructura.

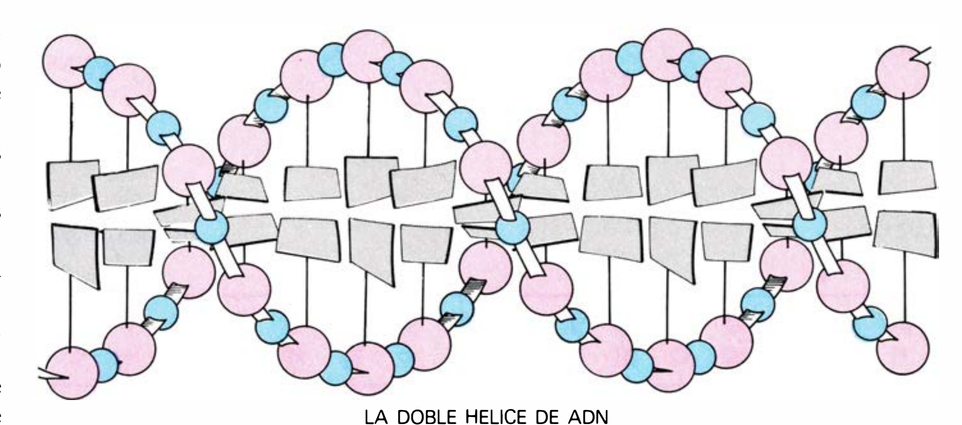
También reaccionan con las bases nitrogenadas los llamados agentes alquilantes. La alquilación consiste en la adición de radicales metilo (CH_3–) o etilo (CH_3–CH_2–) en distintas posiciones según con qué carbono, oxígeno o nitrógeno de la base reaccionen. E igualmente alteran la naturaleza de las bases ciertos derivados de hidrocarburos policíclicos, de aminas aromáticas y los de un número cada vez más elevado de compuestos químicos presentes en el medio ambiente [véase "Tests bacterianos de substancias potencialmente cancerígenas", por Raymond Devoret, INVESTIGACION Y CIENCIA, octubre de 1979. Devoret y el autor han colaborado en varias ocasiones en el descifrado de los procesos de reparación].

Como el perfecto apareamiento de las dos hélices del ADN depende del acoplamiento de las bases, las modificaciones en la estructura de éstas originan distorsiones locales en la conformación de la doble hélice. Estas distorsiones, o cambios en la estructura tridimensional, aunque podrían parecer insignificantes, determinan alteraciones importantes de las funciones biológicas del ADN. Se comprende, entonces, que la restauración de la integridad estructural del ADN, llevada a cabo por los procesos de reparación, sea crucial para la existencia y desarrollo de la célula.

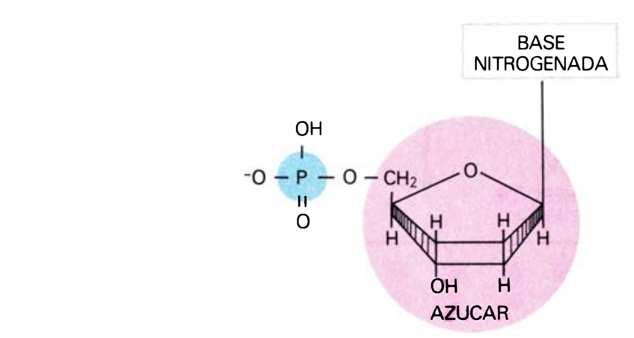
La importancia de la distorsión en la molécula de ADN varía según el tipo de alteración sufrida por las bases componentes. Los dímeros de pirimidina, por ejemplo, dan lugar a una distorsión distinta de la producida por la saturación del anillo de las pirimidinas y de la que resulta de la metilación del oxígeno 6 de la guanina. Los diferentes tipos de distorsiones se repararán por mecanismos diversos que, no obstante, pueden compartir ciertas etapas. Si la reparación de una distorsión empieza, por ejemplo, por la acción de un enzima que, tras detectar la anomalía, actúa sobre ella, esta etapa será realizada por diferentes enzimas, según el tipo de distorsión que haya que reparar, pero de la continuación del proceso pueden encargarse los mismos enzimas. Por otra parte, si diferentes tratamientos originan un mismo tipo de distorsión, ésta será reparada por el mismo mecanismo. Al estudiar los distintos procesos de reparación nos referiremos, principalmente, a los que afectan a las lesiones producidas por la luz ultravioleta, que son los mejor conocidos. Pero ello no debe impedirnos tener una perspectiva más general de dichos procesos de reparación, capaces de actuar sobre todo lo que se reconozca como lesivo para la molécula de ADN, con independencia del agente que la produjo. En la ilustración de la página siguiente se muestran ejemplos de alteraciones en las bases producidas por diversos agentes.

La eliminación de las lesiones del ADN implica la existencia de mecanismos de reconocimiento o de detección de dichas lesiones. Citaremos aquí dos enzimas a quienes compete esa misión detectora. En primer lugar, el enzima fotorreactivante, llamado así porque reconoce los dímeros de pirimidina producidos por la irradiación ultravioleta y, tras acoplarse a ellos y utilizando la energía de la luz (fotoenergía), rompe los enlaces que unían a las dos pirimidinas. Desaparecidas las fuerzas que distorsionaban el ADN, éste recupera su conformación normal.

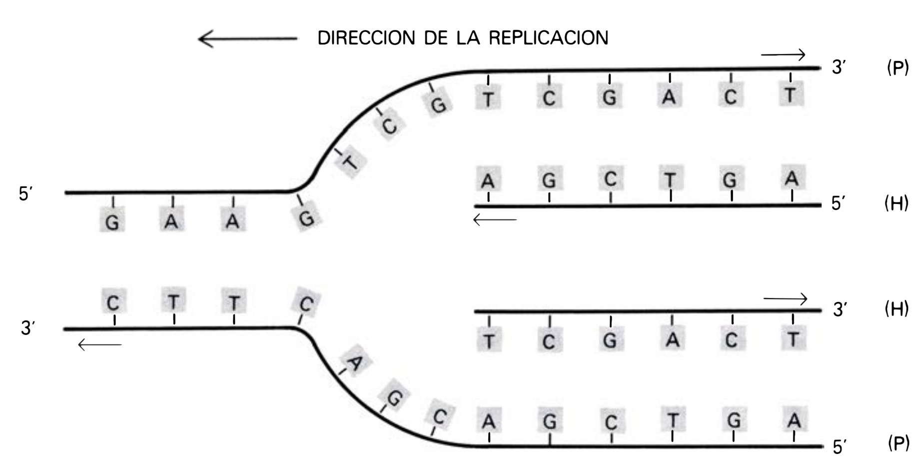
Otro enzima capacitado para la detección de grandes distorsiones es la correndonucleasa-UV, que, además de reconocer la alteración y de acoplarse a ella, realiza un corte en la hélice afectada por la lesión que origina la distorsión (por ejemplo, un dímero de pirimidina). El nombre de correndonucleasa (o endonucleasa de corrección) se debe a que se



LA DOBLE HELICE DE ADN

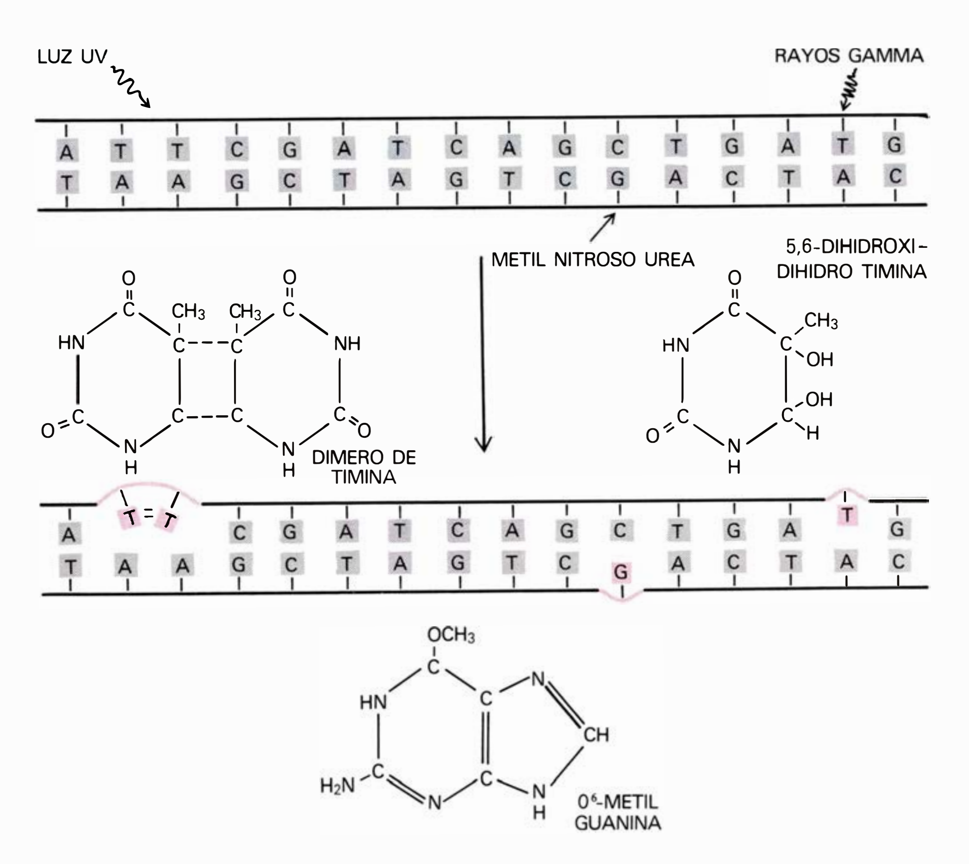


EL NUCLEOTIDO O ESLABON DE LA CADENA



EL MENSAJE GENETICO Y SU REPLICACION

ESTRUCTURA DEL ADN: aparece como una larga cadena compuesta por unidades básicas o eslabones, denominados nucleótidos. Estos consisten en una base nitrogenada, un azúcar (desoxirribosa) y un residuo de ácido fosfórico. La base unida al azúcar mediante un enlace glicosílico constituye un nucleósido. En el ADN hay cuatro bases: adenina (*A*), guanina (*G*), citosina (*C*) y timina (*T*). Durante la duplicación, cada hélice parental (*P*) sirve de molde sobre el que se va sintetizando la hélice hija (*H*).



ALTERACIONES DE LAS BASES, producidas tras los tratamientos con agentes físicos y químicos. Dichas alteraciones originan distorsiones de la doble hélice ADN. En la figura se han señalado en rojo los átomos de cada base implicados en la formación de enlaces hidrógeno con la base complementaria. Las flechas coloreadas indican las posiciones de la molécula de la base susceptibles de reaccionar con los distintos agentes. Por ejemplo, el tratamiento con agentes alquilantes produce la metilación o la etilación de la guanina en el nitrógeno situado en la posición 7 (N–7), en el oxígeno (O–6) unido al carbono 6 o en el grupo amino unido al carbono 2. De esas alteraciones, la metilación del oxígeno 6 tiene probablemente efectos mutagénicos, mientras que la del nitrógeno 7, no; ello puede relacionarse con el hecho de que O–6 pero no N–7 está implicado en la formación del enlace de hidrógeno con la citosina, de modo que la distorsión que resulta entonces es mucho mayor. También se ha demostrado que los derivados de los hidrocarburos aromáticos policíclicos, de gran poder mutagénico, interaccionan con el grupo amino unido al carbono 2 de la guanina, que igualmente interviene en el enlace con la citosina. En la figura se muestra la estructura de un dímero de timina, formado por dos moléculas de esta base que estaban adyacentes en la secuencia del ADN y que, tras absorber la radiación ultravioleta –lo que permite la rotura del doble enlace que une los carbonos 5 y 6–, han reaccionado entre sí. El enlace 5–6 de las pirimidinas reacciona también con los radicales que se forman a consecuencia de un tratamiento con radiaciones ionizantes (así, radicales OH y H formados por radiolisis del agua).

denomina actividad endonucleolítica a la realización de cortes en una o en ambas cadenas de la doble hélice. Por otra parte, se le llama correndonucleasa-UV porque reconoce el tipo de distorsión provocada por los dímeros de pirimidina, principal lesión que resulta de la irradiación con luz ultravioleta (UV), aunque, tal como dijimos, reconoce también las distorsiones del mismo tipo producidas por otros agentes.

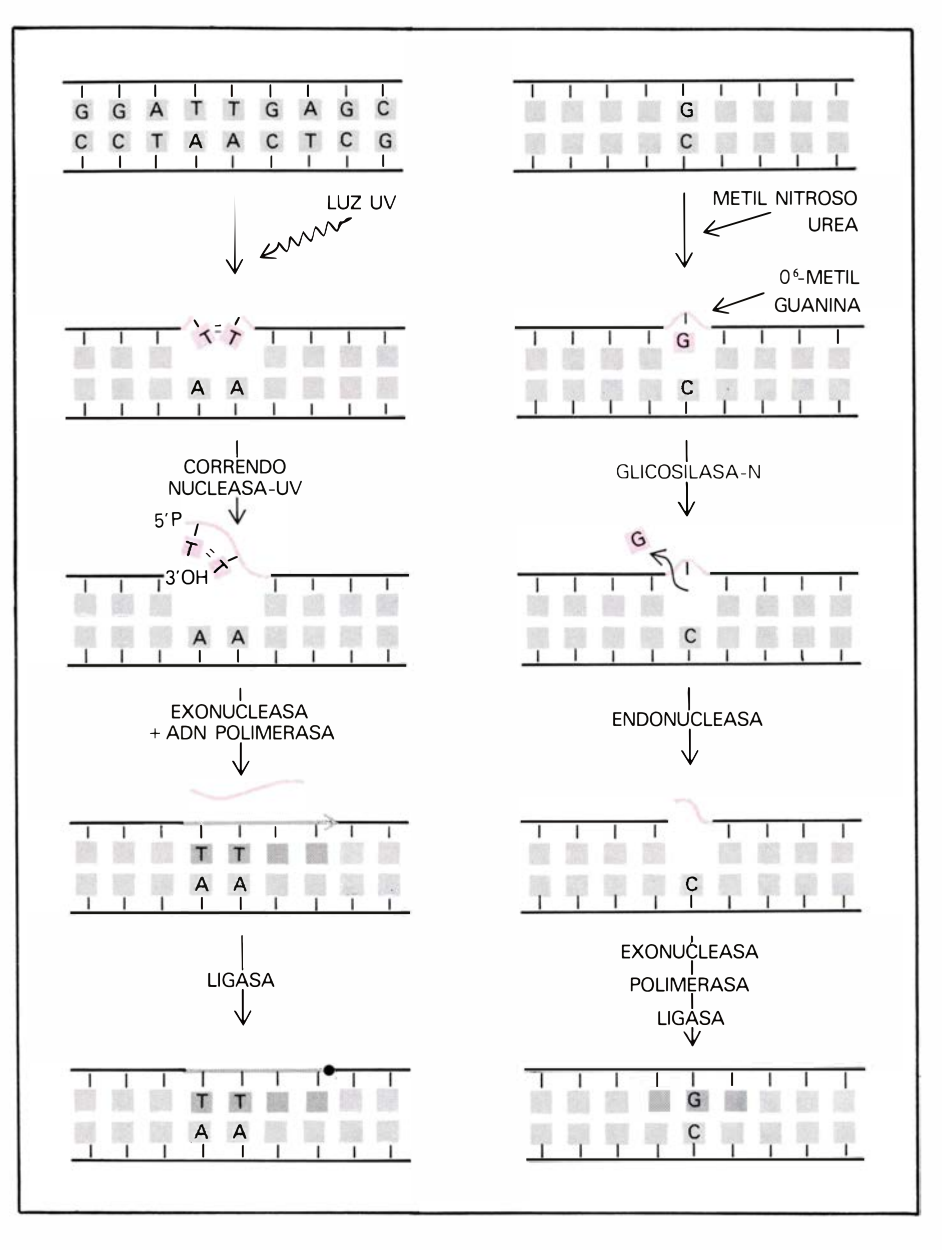
El corte realizado por la correndonucleasa-UV permite la puesta en marcha del proceso de reparación por escisión, tal vez el proceso de reparación más importante de la célula. Este proceso comprende la eliminación del fragmento de ADN lesionado, seguida de la reconstrucción, con exactitud, de dicho fragmento. Esta complicada operación puede ser llevada a cabo por sólo tres enzimas: la citada correndonucleasa-UV (iniciadora del proceso), la ADN-polimerasa I (que elimina y reconstruye el fragmento de ADN) y la ADN-ligasa (que hace la soldadura de los fragmentos nuevo y antiguo, con lo que finaliza el proceso). El corte producido por la correndonucleasa-UV presenta unas características óptimas para que se acople a él la ADN-polimerasa I; ésta elimina, mediante una acción de exonucleasa iniciada en el extremo 5'P del corte, el fragmento de molécula que contiene el dímero. Mediante una actividad de polimerasa, que se inicia en el extremo 3'OH, la ADN-polimerasa I rellena el hueco que dicho enzima, con su otra actividad de exonucleasa, va produciendo. De este modo, la delicada operación de vaciado y rellenado, al ser realizada por un mismo enzima, se sucede con gran celeridad, lo que evita mayores riesgos para la célula. En efecto, una operación de vaciado que dejase durante cierto tiempo un hueco en el ADN, lo expondría al ataque y destrucción por otros enzimas. Tras la acción de la ADN-polimerasa I, y una vez que se han eliminado y añadido unas diez bases, interviene el enzima ADN-ligasa que une el extremo libre del fragmento nuevamente sintetizado con el resto de la cadena, recuperando así el ADN su estructura normal.

Las distorsiones que se producen cuando las bases interaccionan con los agentes alquilantes se reparan por un proceso similar, en parte, al que acabamos de describir. Se inicia con la acción de un enzima, llamado glicosilasa-N, que corta el enlace glicosílico que une la base (concretamente, por uno de sus átomos de nitrógeno) con el resto del

nucleótido. Tras el corte, la base se desplaza del ADN, pero sin que se rompa la continuidad de la cadena de la doble hélice. El hueco dejado por la base es una estructura sensible a la acción de una endonucleasa que hace una incisión en la cadena a la cual estaba unida la base desplazada. Entonces, y tal como ocurría en la reparación por escisión de los dímeros de pirimidina, en la incisión formada actúa, primero, una exonucleasa, que realiza un pequeño hueco, luego, una ADN-polimerasa que lo rellena y, finalmente, una ligasa.

Durante la replicación normal del ADN actuaría un mecanismo análogo al anteriormente descrito. Su misión, entonces, sería la de eliminar, del ADN recién formado, las moléculas de una base pirimidínica, llamada uracilo (U), la cual está presente en otros ácidos nucleicos de la célula (en el ARN), y que, por error, se incorpora en el ADN allí donde debería incorporarse una timina.

Al conocimiento de la existencia de los procesos de reparación se llegó tras el aislamiento de cepas bacterianas (principalmente, de la bacteria *Escherichia coli*) con mutaciones en los genes que codifican los enzimas implicados en dichos procesos. Es decir, cuando se pudo disponer de células que, debido a una mutación en el gen que codifica, por ejemplo, la correndonucleasa-UV o la ADN-polimerasa I, aparecían muy sensibles a la luz ultravioleta y a otros agentes. Estos mutantes, fruto del trabajo paciente del genetista, han constituido el material que ha permitido los estudios a nivel molecular, en los que se basa nuestra comprensión de los procesos reparadores.



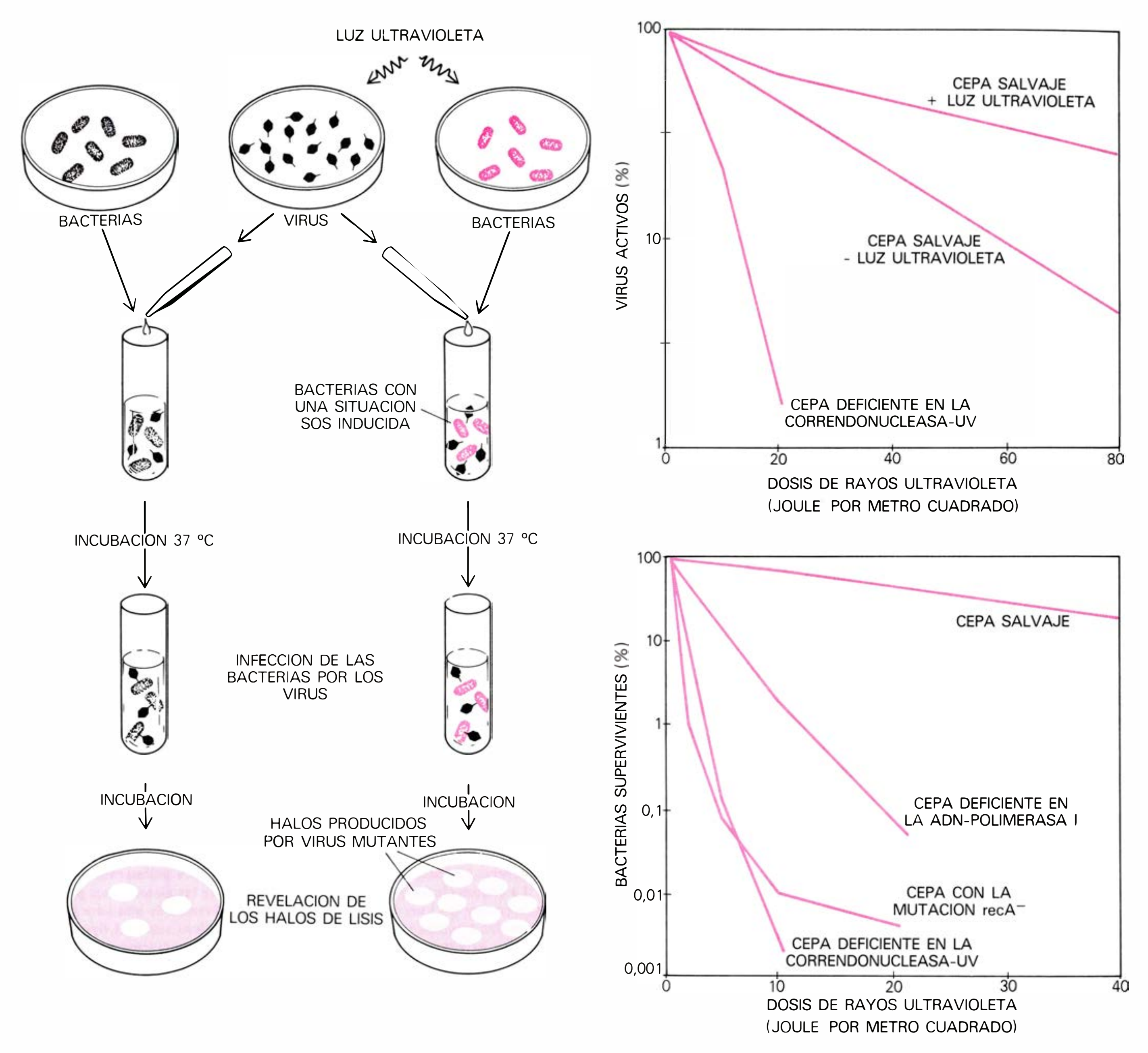
DISTINTAS ETAPAS del proceso de reparación por escisión. Se han representado de forma muy esquemática. En *Escherichia coli*, la actividad de la correndonucleasa-UV depende de la presencia de los productos de tres genes diferentes. Las actividades de exonucleasa y de ADN polimerasa dependen del enzima ADN polimerasa I. La endonucleasa que actúa en la reparación de las distorsiones provocadas por los agentes alquilantes, reconoce los huecos que quedan tras la eliminación de una base por una glicosilasa-N. Estos huecos, llamados también sitios apurínicos o apirimidínicos según el tipo de base que estaba situada en ellos, pueden originarse asimismo por la acción de las radiaciones ionizantes.

La reparación de las lesiones hace que una célula vuelva a ser igual a como era antes del tratamiento que dañó su ADN. Pero podemos preguntarnos si las lesiones no dejan secuelas u originan efectos a largo plazo, es decir, si la reparación neutraliza absolutamente los riesgos que podrían derivarse de las lesiones. Esta cuestión plantea, a su vez, el problema del margen de seguridad que ofrece la reparación ante el ataque de radiaciones o sustancias que alteran el material hereditario.

Si tuviéramos la certeza de que nada le habría de ocurrir a una célula si las lesiones se reparasen, podríamos pensar que, aunque las células de un organismo estuviesen sometidas a una agresión incesante (por ejemplo, por una dosis débil de radiación), como se eliminarían las escasas lesiones producidas, dicha exposición a los agentes nocivos sería inofensiva. Pero podría suceder que la reparación, al borrar las lesiones, tuviese algún efecto negativo; en cuyo caso, nos encontraríamos con que el peligro para la célula dependería de la reparación de dicho daño y no de la propia injuria. Un efecto negativo sería, por ejemplo, que, durante la reparación, se cometiesen errores en el mensaje escrito en el ADN, errores que originarían mutaciones. Hay pruebas de que tanto la fotorreactivación como la escisión de lesiones son procesos que no introducen errores en el ADN. Sin embargo, cuando en el proceso de escisión las operaciones de vaciado y rellenado se realizan por otros enzimas distintos de la ADN-polimerasa I (otras exonucleasas y ADN-polimerasas), podrían originarse un número importante de errores en el ADN. Resumiendo, los procesos de reparación que eliminan las lesiones son neutralizadores eficaces del daño ocasionado en el material genético de la célula.

De la presencia de lesiones en el ADN sí pueden derivarse consecuencias fatales a largo plazo cuando se da la circunstancia de que, antes de la eliminación completa de las lesiones, el ADN realiza una de sus funciones más importantes: la de replicarse. Este hecho plantea la cuestión de si el ADN que muestra

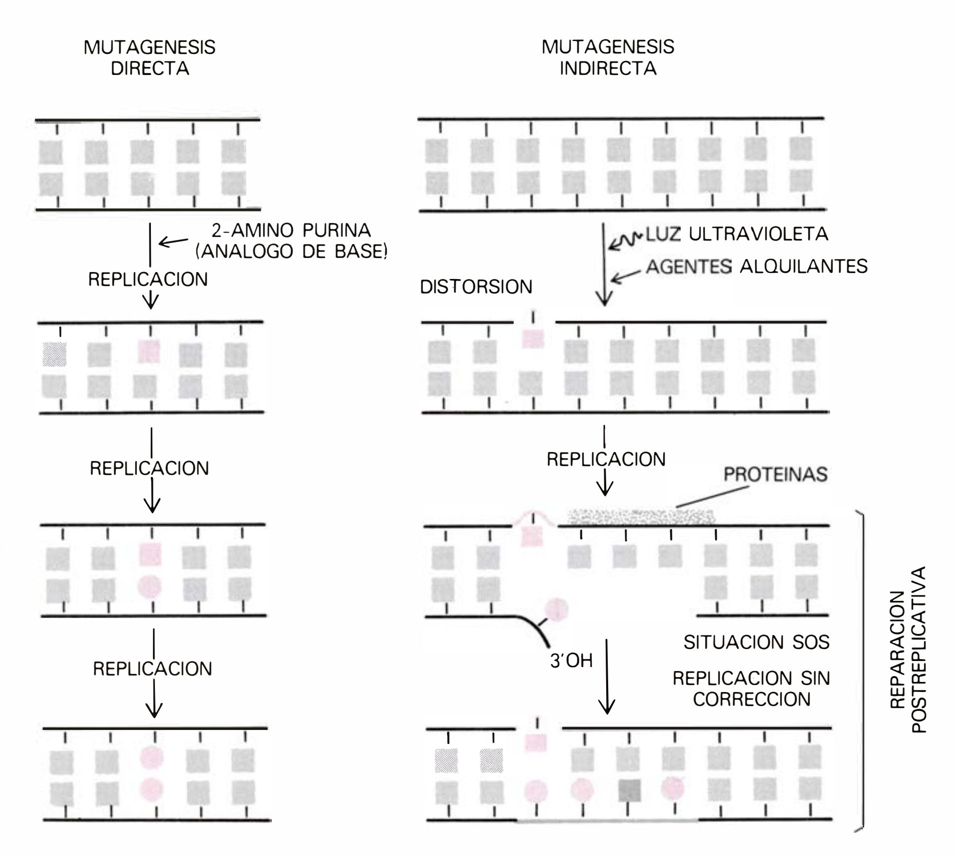
grandes distorsiones en su conformación puede servir o no de molde para la síntesis de nuevo ADN. Los resultados experimentales demuestran: 1) que un ADN lesionado puede replicarse; 2) que los fragmentos de ADN sintetizados a partir de un molde con la estructura alterada son, inicialmente, de tamaño más pequeño que los sintetizados a partir de un molde intacto y, posteriormente, alcanzan la talla normal. El modelo generalmente aceptado de lo que ocurre implica la formación, frente a las distorsiones, de discontinuidades en la nueva cadena de ADN, discontinuidades que se van rellenando con el transcurso del tiempo. A los procesos responsables de que el ADN lesionado sirva de molde, en el que se forman discontinuidades, y al posterior rellenado de éstas, se les engloba bajo la denominación de repara-



EN LA EXPERIENCIA DE WEIGLE se estudia la inactivación, o pérdida del poder citopático, de un virus. Para ello, se infectan bacterias con virus que han sido irradiados con luz ultravioleta. Tras incubar a 37 grados Celsius durante un corto período de tiempo para facilitar la adsorción de las partículas víricas a los receptores de la membrana bacteriana, se vierte la mezcla de bacterias y virus sobre una placa de Petri que contiene un medio de cultivo sólido. A la citada mezcla, antes de ser vertida, se le añade un pequeño volumen de agar poco concentrado (blando) mantenido líquido a 48 grados C. Se forma así, sobre el medio sólido de base, una pequeña capa semifluida, que se solidifica rápidamente por enfriamiento, y en la que las células van a crecer de manera confluente formando lo que podríamos denominar un tapiz de bacterias. Allí donde haya caído una bacteria infectada por un virus, si éste es activo, tras multiplicarse en la célula la lisará y se liberarán entonces un gran número de virus que infectarán a las células próximas, lisándolas también, y así sucesivamente hasta que, al perder humedad la capa de agar, los virus no puedan propagarse e infectar nuevas bacterias. El centro infeccioso origina, en consecuencia, una zona o halo de lisis en el tapiz de crecimiento bacteriano confluente. Cuando las bacterias que se infectan han sido, a su vez, irradiadas con luz ultravioleta, se obtiene un incremento en la frecuencia de virus mutantes y un mayor número de halos de lisis, lo que indica una menor inactivación de los virus. Los resultados de una experiencia de Weigle, realizada con la bacteria *Escherichia coli* y el bacteriófago *lambda*, se muestran en la figura. Aparece también la curva de inactivación de un virus en una cepa mutante que es deficiente en la reparación por escisión por carecer de la actividad del enzima correndonucleasa-UV. La reparación de un virus depende generalmente de funciones celulares, aunque a veces, los virus poseen funciones para su propia reparación. Los efectos letales de un tratamiento que lesiona el ADN pueden estudiarse también utilizando bacterias. En este caso, se analiza la muerte celular, es decir la pérdida de la capacidad de división de una célula, lo que se traduce en la imposibilidad de que origine una colonia al ser sembrada sobre un medio de cultivo sólido. Los resultados muestran las curvas de supervivencia de cepas de la bacteria *Escherichia coli* que poseen mutaciones en genes que codifican proteínas implicadas en procesos de reparación.

ción post-replicativa. La importancia de esta reparación post-replicativa radica, por una parte, en que de ella depende la tolerancia de las lesiones que persisten en el ADN, y por otra, en que, como veremos más adelante, a ella aparecen vinculadas las consecuencias que, a largo plazo, se derivan de los tratamientos que lesionan el ADN.

La tolerancia celular para con las lesiones sufridas comprendería varias etapas que, como las de cualquier secuencia metabólica, dependerían de la acción de ciertas proteínas; por tanto, podrían hallarse mutantes deficientes en los mecanismos de tolerancia de lesiones (es decir, que no lograsen sobrepasarlas). Tales mutantes se han aislado ya en *Escherichia coli.* Se trata de mutantes en el gen llamado *recA* (las tres letras, *rec,* se deben a que este gen tiene un importante papel en la recombinación genética; dichos mutantes fueron aislados, precisamente, por su deficiencia en la recombinación). Los mutantes en el gen *recA* son sensibles a la luz ultravioleta y demás agentes que lesionan el ADN. Ello no impide que lleven a buen término la escisión de lesiones. En ellos, el ADN sintetizado a partir de un molde que contiene alteraciones no alcanza la talla normal, sino que permanece en pequeños fragmentos. Estos resultados indican que la mutación *recA* afecta a la reparación post-replicativa.



LA REPARACION POST-REPLICATIVA aparece relacionada con la mutagénesis. Esta puede ser el resultado de la incorporación del análogo de una base durante la replicación del ADN (por ejemplo, la 2-aminopurina, parecida a la adenina, una vez que ha sido incorporada en el ADN puede acoplarse con la citosina; la secuencia de hechos sería: A = T→2 − AP = T→2 − AP = C→ G = C, es decir, un par A = T se ha convertido en uno G = C). A esta mutagénesis la llamamos directa. La mutagénesis indirecta se inicia con una lesión en el ADN (lesión premutagénica); aparece vinculada a la serie de etapas que constituyen la reparación post-replicativa, la cual comprendería la formación de una discontinuidad o fragmento de ADN de simple hélice, y su posterior rellenado. Dicho ADN de simple hélice sería el sustrato de proteínas que se acoplarían a él y lo mantendrían en una configuración adecuada para servir de molde durante una polimerización. Si este molde posee una anomalía estructural importante (por ejemplo, un dímero de pirimidina o una base con una alteración tal que le impida acoplarse con ninguna otra), toda base que se incorpore enfrente de la anomalía será reconocida como incorrecta y eliminada por la exonucleasa 3'-5'. Hay pruebas de que se produce entonces una importante hidrólisis de nucleósidos trifosfato, que son incorporados y eliminados, del ADN, lo que origina un incremento de nucleósidos monofosfato; algunos resultados indican que, a consecuencia de dicho incremento, podría inhibirse la actividad correctora de la ADN-polimerasa. Las mutaciones producidas por mutagénesis indirecta estarán localizadas no sólo frente a la distorsión, sino en cualquier fragmento de ácido desoxirribonucleico que se esté replicando mientras persista en la célula una situación SOS (situación, entre otras cosas, de inhibición de la actividad correctora de la ADN-polimerasa), aun cuando dicho fragmento no posea ninguna lesión. Una situación SOS originaría una mutagénesis generalizada.

El hecho de que estos mutantes deficientes en la reparación post-replicativa lo sean, también, en la recombinación, ha dado pie a hipótesis que sugieren la intervención de un mecanismo de recombinación en dicha reparación. Que la recombinación pueda estar implicada en la reparación se ha venido sospechando desde mucho antes de que se descubriese la existencia de procesos enzimáticos de reparación. En efecto, desde hace treinta años se sabe que, tras infectar una bacteria con varios virus lesionados con luz ultravioleta, se producen intercambios de fragmentos entre las moléculas lesionadas de ADN de los distintos virus, pudiendo originarse, a consecuencia de ello, una molécula intacta del ADN vírico. [La experiencia ha sido descrita por M. Delbruck y M. B. Delbruck en *Scientific American,* noviembre de 1948.]

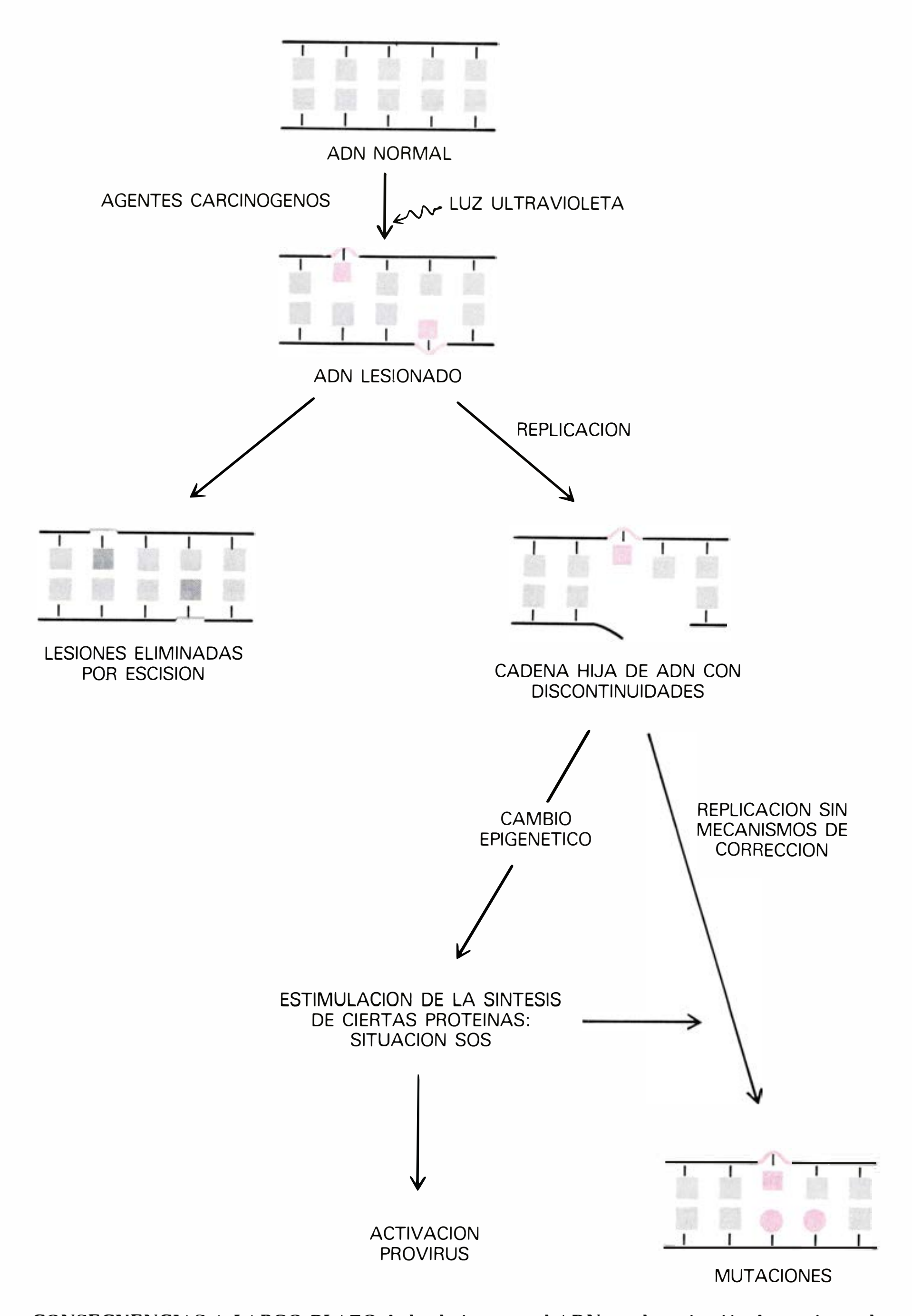
Cuando un fragmento de ADN lesionado se encuentre en presencia de otro fragmento homólogo (con la misma secuencia de bases), es fácil suponer que la recombinación desempeñará un papel como mecanismo de reparación. Este posible papel adquirió más importancia cuando, en 1968, Paul Howard-Flanders propuso la hipótesis de que el mecanismo fundamental de la reparación post-replicativa sería una recombinación entre las dos dobles hélices hijas. A consecuencia de esto, en muchas ocasiones se ha considerado que la reparación post-replicativa era equivalente a la reparación por recombinación. Sin embargo, y aunque la recombinación podría estar implicada en parte de la tolerancia de lesiones, cabe tener en cuenta la posibilidad de que dicha tolerancia dependa también de otros mecanismos distintos del de recombinación genética.

La supervivencia de una célula después de un tratamiento que lesione su ADN dependerá de la eliminación de las injurias sufridas, antes de que el ADN se replique (eliminación que tendrá lugar, fundamentalmente, mediante el proceso de escisión), y de la capacidad de dicha célula para tolerar las lesiones que queden en su ADN cuando éste se replique. Si no existiese esta capacidad de tolerancia, una sola lesión en la larga molécula de ADN sería suficiente para originar la muerte de la célula. Esto ha podido comprobarse experimentalmente en mutantes de *E. coli* deficientes en la escisión de dímeros y en la tolerancia de lesiones.

Si irradiamos con luz ultravioleta un cultivo bacteriano, obtendremos tres subpoblaciones celulares: *a)* células que han reparado completamente las lesiones producidas en su ADN y en las que, por tanto, no ha lugar para que funcionen los mecanismos de tolerancia de le-

siones; *b*) células que toleran la totalidad de las lesiones residuales, y *c*) células en las cuales hay lesiones residuales que no se toleran y que, por consiguiente, mueren. En esta situación, podría conseguirse un incremento de la viabilidad aumentando la eficacia de la reparación por escisión, lo que se lograría alargando el intervalo de tiempo que media entre el momento del tratamiento y el instante en que la zona lesionada ha de replicarse. También se incrementaría la viabilidad si se pudiese mejorar el rendimiento de los mecanismos encargados de la reparación post-replicativa.

Veamos cómo es posible mejorar la capacidad reparadora de una célula. Jean Weigle, de la Universidad de California, fue el primero en realizar un experimento que confirmaba el supuesto anterior. En 1953, Weigle quería obtener mutantes del virus bacteriano, o bacteriófago, *lambda* (el mejor conocido entre todos los bacteriófagos que infectan a *E. coli*), y decidió utilizar la luz ultravioleta como agente mutágeno. Irradió la suspensión de virus y, con las partículas víricas lesionadas, infectó bacterias de *E. coli*, con la esperanza de hallar mutantes entre la descendencia de sus virus irradiados. Sin embargo, dichos mutantes no aparecieron. Este resultado hizo suponer que la presencia de lesiones en el ADN no era suficiente para desencadenar mutaciones, sino que hacía falta algo más.



CONSECUENCIAS A LARGO PLAZO de las lesiones en el ADN son la activación de provirus y la producción de mutaciones. Las lesiones eliminadas por escisión no intervienen en dichas consecuencias. Las lesiones no eliminadas, por el contrario, al interferir con la replicación, provocan cambios epigenéticos en la célula (por ejemplo, la estimulación de la síntesis de la proteína *recA*) y la célula pasa a encontrarse en "situación SOS", vista antes. Esta situación, a través de mecanismos todavía no bien conocidos, provoca la activación de provirus latentes (en estado silencioso) en la célula y la mutagénesis.

Weigle infectó luego con sus fagos irradiados bacterias que, a su vez, habían sido irradiadas con una dosis débil de luz ultravioleta antes de la infección. Obtuvo entonces un gran número de mutantes del virus. Pero, además, observó que el número de fagos activos recuperados después de la infección se había incrementado mucho. Era como si, al estar la bacteria irradiada, hubiese habido una mejor reparación del ADN vírico lesionado. Podía hipotetizarse que los mutantes aparecían entre la población suplementaria de fagos reparados, lo que equivalía a admitir que estaba funcionando un proceso que mejoraba la reparación del ADN pero que, al mismo tiempo, originaba un incremento de la mutagénesis. Se llama efecto Weigle (en honor de Jean Weigle, quien lo observó por vez primera) al aumento de la reparación del virus asociado con un incremento de las mutaciones.

¿Qué mecanismo de reparación es el responsable del efecto Weigle? Durante más de veinte años se han venido trabajando en diversos laboratorios intentando contestar ese interrogante. Al principio se pensó que resultaría de alguno de los procesos ya conocidos, por ejemplo la reparación por escisión, o que dependería de un mecanismo de recombinación. La primera posibilidad fue descartada al demostrarse que este efecto ocurría en los mutantes deficientes en la escisión. En cuanto a la segunda posibilidad, encontró un apoyo al verse que los mutantes en el gen *recA* eran deficientes en la recombinación y en el efecto de Weigle (en ellos no se producía dicho efecto). Sin embargo, se demostró que el incremento de fagos reparados no se correlacionaba con un incremento en el número de recombinantes, como hubiese sido de esperar si la reparación se debiera al proceso recombinante.

Se pensó, entonces, que el efecto Weigle podría estar relacionado con la tolerancia para con lesiones del ADN. Por otra parte, al estudiar los mutantes deficientes en el efecto Weigle, se observó que tales mutantes eran muy sensibles a las exposiciones que alteraban la estruc-

tura del ADN y, además, que esos tratamientos no originaban mutaciones. Todo indicaba que en estos mutantes faltaba un tipo de reparación que era el que conduciría a la mutagénesis. Podía suponerse, entonces, que el efecto Weigle no sólo actuaba en la reparación de un ADN vírico lesionado, sino que intervenía también en la reparación del ADN celular. A consecuencia de la reparación detectada en la experiencia de Weigle, se originarían precisamente los mutantes de la célula que se encontraban tras la aplicación de un tratamiento mutagénico. Y, según la hipótesis de que el efecto Weigle se produciría durante la tolerancia de lesiones, el fenómeno de la mutagénesis aparecería vinculado a dicha tolerancia. Tal como vimos anteriormente, parte del mecanismo de tolerancia implicaría el rellenado de las discontinuidades que existirían en una hélice de ADN, situadas frente a las lesiones no eliminadas. Podía pensarse que, durante la operación de rellenado, llevada a cabo por una ADN-polimerasa, se cometerían los errores que darían origen a las mutaciones.

TRATAMIENTO	EJEMPLOS DE ALTERACION DE LAS BASES	CONSECUENCIAS DE LA ALTERACION
Radiaciones no ionizantes: luz ultravioleta	Dímeros de pirimidina	Gran distorsión reconocida por la correndonucleasa-UV
Radiaciones ionizantes: rayos gamma	Hidroxiperoxidación en el doble enlace (5-6) de las pirimidinas	Pequeña distorsión reconocida por la endonucleasa-gamma
Agentes alquilantes: metil nitroso urea, etil nitroso urea	Metilación o etilación del oxígeno 6 de la guanina	Gran distorsión reconocida por una glicosilasa-N y por una endonucleasa
Derivados de los hidrocarburos aromáticos policíclicos: dihidrodiol benzo (a) pireno	Formación de un enlace con el grupo amino unido al nitrógeno 2 de la guanina	Distorsiones reparadas por mecanismos no del todo conocidos
Aminas aromáticas: 2 acetil aminofluoreno (AAF)	Formación de un enlace con el carbono 8 de la guanina	Distorsiones reparadas por mecanismos no del todo conocidos

LOS DISTINTOS CARCINOGENOS, radiaciones y agentes químicos, provocan diferentes alteraciones de las bases que distorsionan en mayor o menor grado la doble hélice de ADN. Un agente dado puede originar diversas alteraciones. Se indican las que originan mayores consecuencias biológicas.

Pero sale al paso entonces una objeción: existen unos mecanismos de seguridad en la replicación del ADN que hacen casi imposible que una ADN-polimerasa cometa errores. En efecto, característica fundamental de la duplicación del ADN es su gran fidelidad. Podemos decir que los mecanismos de replicación, no sólo copian las hélices del ADN, sino que las copian sin equivocarse (algunos cálculos indican que se cometería un error por cada diez mil millones de pares de bases duplicados).

Este elevado grado de precisión se debe a la acción de un mecanismo corrector que, una vez añadida una base, si ésta no es la que corresponde (o sea, la complementaria a la base de la hélice que sirve de molde), la desecha. Esta operación depende de la llamada actividad correctora de pruebas, realizada por una exonucleasa que actúa en la dirección 3'-5'. La propia ADN-polimerasa cumple esta función de exonucleasa; un solo enzima, pues, va sacando una réplica de la secuencia de bases y, al mismo tiempo, verificando que la réplica sea correcta. La replicación sólo puede avanzar si la última base añadida es la adecuada, de modo que si se añadiesen siempre bases erróneas, la replicación quedaría bloqueada. La polimerización en presencia de una lesión equivaldría a una situación de mal apareamiento, encontrándonos, entonces, con una serie inacabada de polimerización-corrección. La polimerización en un molde anormal sólo sería posible si no funcionase el mecanismo corrector, lo que supondría incrementar de manera significativa el riesgo de errores.

La hipótesis de que el efecto Weigle resultaría de una polimerización errónea fue propuesta, en 1974, por Miroslav Radman, quien llamó a esta replicación "replicación SOS" o "replicación de emergencia" (posteriormente, Evelyn Witkin la rebautizó como "reparación SOS"). La hipótesis de Radman predecía que la reparación SOS no estaría funcionando siempre en la célula, sino que sería un proceso activable por los tratamientos que lesionan el ADN.

Una experiencia que apoya la hipótesis de que la reparación SOS se basa en el funcionamiento de un mecanismo que permitiría la polimerización en un molde de ADN lesionado ha sido realizada por Radman y sus colaboradores de la Universidad Libre de Bruselas. Para ello utilizaron el bacteriófago llamado ØX 174, cuyo material genético consiste en una molécula de ADN con una sola hélice (en lugar de las dos que tiene el ADN de todos los seres vivos). Cuando el ADN de dicho fago contiene lesiones producidas por la irradiación con luz ultravioleta, no puede ser replicado ya que la síntesis de ADN cesa en cuanto hay que copiar el fragmento lesionado (en este caso, a diferencia de lo que ocurre cuando el ADN tiene dos hélices, no se originan discontinuidades en presencia de lesiones, sino que éstas bloquean absolutamente la replicación de dicho ácido nucleico.

Radman irradió con luz ultravioleta una suspensión del fago ØX 174 e infectó, con lo fagos lesionados, bacterias de *E. coli* que habían sido irradiadas o no, también con luz ultravioleta, antes de la infección. Analizó, entonces, el fragmento de ADN vírico que se había replicado en las bacterias irradiadas y en las no irradiadas, encontrando que, en las primeras, las lesiones presentes en el ADN vírico no habían constituido un bloqueo para la replicación, habiéndose copiado en ellas el ADN que poseía distorsiones. Este resultado, por otra parte, podía correlacionarse con el elevado número de mutantes que se encontraban en la descendencia de los virus que infectaron las bacterias irradiadas.

¿Qué enzima realizaría la replicación de un ADN lesionado? En *E. coli* se conocen tres ADN-polimerasas con actividad correctora y, por tanto, no podrían realizar la polimerización a no ser que dicha actividad correctora se inhibiera transitoriamente. Cabe, también, la posibilidad, menos probable, de que actuase una ADN-polimerasa todavía desconocida.

La reparación SOS es dependiente del producto del gen llamado *recA*, recientemente aislado. Se trata de una proteína (a la que se conocía desde hace varios años con el nombre de proteína *X*), que puede acoplarse al ADN y cuya función precisa todavía se desconoce. La síntesis de esta proteína parece estar sometida a una compleja regulación. Se ha visto que, después de aplicar a una bacteria un tratamiento que lesiona el ADN, la síntesis de la proteína *recA* experimenta una gran estimulación, llegando a constituir hasta el uno por ciento del total de las proteínas de la célula. Recordando que dichos tratamientos desencadenan la reparación SOS puede establecerse un

paralelismo entre el carácter activable de este efecto y la estimulación de la síntesis de la proteína *recA*.

Resumiendo, podemos decir que hoy se dispone de pruebas experimentales de que la reparación SOS es, en parte, responsable de la tolerancia de lesiones no reparadas, tolerancia que sería una consecuencia de la posibilidad de replicar, cometiendo errores, un molde de ADN lesionado y que requeriría la estimulación de la síntesis de la proteína *recA*.

El carácter activable de la reparación SOS es comprensible, pues su expresión, de un modo constitutivo, llevaría a un gran nivel de mutagénesis, con los consiguientes riesgos que se derivarían de una alteración importante del genotipo. Lo que nos interesa destacar es que la aparición de la actividad mutagénica viene determinada por la presencia, en el ADN, de unas lesiones que deben ser toleradas para que la célula sobreviva. La situación mutagénica tiene un carácter extraordinario y es una respuesta de la célula a una agresión a su ADN, respuesta que desencadenan las propias lesiones del ADN.

En la elaboración de la hipótesis de la reparación SOS como fenómeno activable en la célula desempeñó un papel importante el hecho de que las condiciones genéticas y fisiológicas que hacían posible dicha reparación eran exactamente las mismas que se requerían para la inducción o activación de un provirus. Para explicar qué es un provirus nos es necesario referirnos a un tipo particular de relación que puede establecerse entre los materiales genéticos de una célula y de un virus. Hay virus que, tras infectar una célula, se desarrollan en ella y acaban por destruirla. Existe, sin embargo, otro tipo de virus cuyo material genético, después de entrar en la célula, se inserta en el material genético celular y se transmite a las células hijas junto con los demás genes de la célula parental. Al material genético del virus, integrado en el de la célula, se le llama provirus.

Los genes del provirus no expresan su mensaje; si lo hiciesen, tendría lugar el desarrollo vírico que causaría la muerte de la célula. En *E. coli* se ha visto que el estado silencioso de los genes del provirus se mantiene merced a la acción de una proteína, llamada represor, que está codificada por el único gen activo del provirus. El represor puede inactivarse si se trata la célula con agentes que lesionan el ADN; esta inactivación es la causa de la inducción del provirus, fenómeno descubierto, en 1950, por André Lwoff en el Instituto Pasteur. Como dijimos, la inducción del provirus depende de las mismas condiciones que la reparación SOS (presencia de lesiones en el ADN de la célula, alto nivel de proteína *recA*) y aparece, por tanto, como una consecuencia del desequilibrio celular provocado por las alteraciones del ADN. Dicho desequilibrio provoca también otros fenómenos celulares (tales como el bloqueo de la división celular o de la respiración). Se ha avanzado la hipótesis de que estos fenómenos dependerían de la activación de unas funciones a las que se engloba bajo la denominación de funciones SOS. Normalmente, estas funciones no se expresarían, como dijimos ocurriría con la reparación SOS, pues lo impedirían proteínas represoras.

¿Cuál es el mecanismo por el que se activan las funciones SOS en respuesta a unas condiciones determinadas? Jeffrey y Christine Roberts, trabajando en la Universidad de Harvard, han demostrado que, en *E. coli*, tras la aplicación de un tratamiento inductor, se destruye el represor del provirus por una proteasa (una proteína capaz de cortar en fragmentos a otras proteínas). Posteriormente se han obtenido pruebas experimentales de que, si después de aplicar el tratamiento inductor, se incubaban las bacterias en presencia de sustancias que inhibían la acción de las proteasas, no se observaba la inducción de las diversas funciones SOS (no se inducía el provirus, no se producían mutaciones, ni se bloqueaba la división celular). Estos resultados apoyan la hipótesis de que existe un sistema común de regulación de las funciones SOS, si bien los mecanismos precisos de esta regulación son todavía desconocidos y el papel desempeñado en ella por las proteasas es tema de especulación.

En la ilustración de la página 12, presentamos un esquema de la secuencia de hechos que, partiendo de las lesiones en el ADN, conducen a fenómenos tales como cambios en la expresión de genes (estimulación de la síntesis de la proteína *recA*), modificaciones de las actividades de enzimas (bloqueo de la actividad correctora de las ADN-polimerasas), incremento en el nivel de mutaciones o activación de los genes de un provirus.

Dicho esquema sirve de respuesta a la cuestión sobre los efectos que pueden derivarse de los tratamientos que lesionan el ADN. Por otra parte, nos muestra el avance realizado en el conocimiento de los mecanismos moleculares que conducen a tales efectos. Ahora, no sólo podemos decir que la irradiación con luz ultravioleta produce mutaciones, sino que podemos explicar, en gran parte, los mecanismos de la propia célula a través de los cuales, partiendo de la lesión que la radiación ultravioleta ha originado en el ADN, se llega a un cambio estable y hereditario en el mensaje genético de dicha célula, es decir, a una mutación. Y lo mismo que decimos de los efectos de la luz ultravioleta, lo podríamos decir de los de otras radiaciones y agentes químicos.

Los conceptos que hemos presentado a lo largo de este artículo cobran un nuevo valor a la luz del hecho, cada vez más evidente, de que los agentes carcinogénicos originarían el cáncer a través de las lesiones que causan en el ADN, es decir, provocando lo que podríamos denominar una toxicidad genética. Tal posibilidad supone que los mecanismos de reparación desempeñarían un importante papel eliminando lesiones potencialmente carcinogénicas y haciendo que la agresión por dichos agentes no fuese un hecho irremediable.

Los estudios realizados con animales superiores indican que la probabilidad de que se origine un cáncer de un órgano determinado es mayor cuanto menor es su capacidad para reparar las lesiones en el ADN. Concretamente, en el hombre, se conoce una enfermedad hereditaria, llamada *Xeroderma pigmentosum*, a consecuencia de la cual las lesiones producidas en el ADN de los fibroblastos de la piel por la radiación ultravioleta, presente en la luz solar, no son eliminadas, y ello conduce a carcinomas de la piel. Se ha podido demostrar que las células de las personas afectadas por esta enfermedad presentan, a nivel molecular, anomalías similares a las de las bacterias deficientes en la correndocleasa, enzima que reconoce las lesiones producidas por la luz ultravioleta y hace posible la reparación por escisión.

Otros estudios, realizados en ratas, han mostrado que la etil-nitrosourea provoca solamente tumores en el sistema nervioso. La etil-nitrosourea produce la etilación del oxígeno situado en la posición 6 de la guanina, y se ha podido comprobar que, si bien la etilación tiene lugar en los diversos tejidos del animal, dicha lesión está todavía presente en el cerebro diez días después del tratamiento, mientras que ya no es detectable en el hígado, es decir, ha sido reparada. Un resultado similar es el obtenido, también en ratas, con la dimetilnitrosamina, que produce tumores en el riñón, donde las lesiones no son elimi-

nadas, pero no en el hígado, donde actúa una eficiente reparación.

¿A través de qué mecanismos celulares las lesiones no reparadas conducen a la carcinogénesis? Este problema apasionante queda planteado y es de esperar que, poco a poco, se avanzará en su resolución. [Véase "El problema del cáncer", por John Cairns, INVESTIGACION Y CIENCIA, octubre, 1976.] Recordemos que, de las lesiones que no se eliminan antes de la replicación del ADN, pueden derivarse cambios, los cuales, podríamos decir, acompañan al esfuerzo que la célula ha de hacer para sobrevivir, y que se manifiestan en fenómenos como el de la mutagénesis. Cabe pensar que el mecanismo que lleva a dicha mutagénesis esté relacionado, de alguna manera, con el que llevaría a la carcinogénesis.

La posible relación entre mutagénesis y carcinogénesis ha sido explotada para poner a punto sistemas experimentales en los que se mide la capacidad que tienen diferentes agentes para producir mutaciones en bacterias. Si un agente dado tiene alto poder mutagénico, hay una gran probabilidad de que sea carcinogénico. También se han puesto a punto otros sistemas en los que se analiza el poder inductor del provirus de los diferentes agentes. Tales sistemas se están revelando muy útiles y han respondido a la necesidad que se tenía de disponer de métodos rápidos y económicos para evaluar lo que hemos llamado toxicidad genética de agentes potencialmente carcinogénicos presentes en el medio ambiente.

La correlación que parece existir entre la potencia mutagénica y la carcinogénica de un determinado agente se ha interpretado, en algunos casos, para defender la teoría que considera las mutaciones que ocurren en células somáticas como la causa del cáncer. Creemos, sin embargo, que dicha correlación no excluye que haya una parte de verdad en otras teorías que presentan el cáncer como una consecuencia de cambios epigenéticos (cambios en la expresión de genes) o de la acción de virus oncogénicos. Si recordamos el esquema resumen que expusimos sobre las consecuencias de la replicación de un ADN lesionado, vemos que en un organismo simple como es la bacteria *E. coli*, la mutagénesis, la estimulación de genes y la activación de provirus aparecen relacionados y dependientes de una actividad en la que el ADN ocupa un lugar central. El mismo que, seguramente, ocupa en los procesos complejos que hacen posible la diferenciación celular y la evolución.

Causas de la diabetes

Hay dos tipos fundamentales de diabetes: una forma de comienzo juvenil y otra de comienzo en la madurez. La forma de inicio juvenil parece desarrollarse a partir de una compleja interacción entre factores genéticos y ambientales

Abner Louis Notkins

La diabetes mellitus y sus complicaciones constituyen la tercera causa de muerte en los Estados Unido precedida sólo por las enfermedades cardiovasculares y el cáncer. De acuerdo con un informe publicado por la Comisión Nacional de Diabetes en 1976, 10 millones de americanos, cerca del cinco por ciento de la población, pueden tener diabetes, incidencia que aumenta cada año. Los efectos directos e indirectos de la diabetes en la economía estadounidense son enormes, sobrepasando los 5000 millones de dólares por año. Si la tendencia actual continúa, el norteamericano que venga hoy al mundo tendrá una probabilidad de un 20 por ciento de desarrollar, andando el tiempo, dicha enfermedad. La probabilidad de convertirse en diabético se duplica cada diez años de la vida del individuo y con cada 20 por ciento de exceso de su peso corporal.

Por otra parte, aunque la sintomatología aguda, y a menudo letal, de la diabetes puede controlarse a través de la terapéutica insulínica, las complicaciones a largo plazo de la enfermedad reducen la esperanza de vida en un tercio. Comparados con los no diabéticos, los diabéticos presentan un porcentaje de ceguera 25 veces mayor, 17 veces mayor para las enfermedades renales, cinco veces para la gangrena y dos para las cardiopatías.

Muchos aspectos de la diabetes permanecen todavía envueltos por el halo del misterio. Sin embargo, trabajos recientes en tres frentes sin aparente relación –genética, inmunología y virología– han coincidido en la opinión de que la diabetes constituye un grupo heterogéneo de enfermedades más que una sola entidad. Estos estudios señalan también que la diabetes es el resultado final de una compleja interacción entre los factores genéticos individuales y los ambientales.

La diabetes mellitus es una enfermedad conocida desde la antigüedad. La primera descripción de sus síntomas se encuentra en el papiro de Ebers (Egipto), que data del año 1500 a. C. En el siglo II d. C., Areteo de Capadocia la denominó diabetes, palabra griega que significa "correr a través de un sifón". "La diabetes", escribe, "es un extraño mal que consiste en que la carne y los huesos fluyen juntos por la orina". Esta es una descripción imaginativa de los más notables síntomas de la diabetes: un gran flujo de orina acompañado por una sed extrema y gran apetito que, sin embargo, resultan en el desgaste de músculos y grasa, y que a menudo termina en coma y muerte. En el siglo VI los médicos hindúes se dieron cuenta de que la orina de los diabéticos tenía sabor dulce. Pero hasta el siglo XVIII el sabor dulce de la orina no se identificó como glucosa, añadiéndose la palabra mellitus: "sabor a miel".

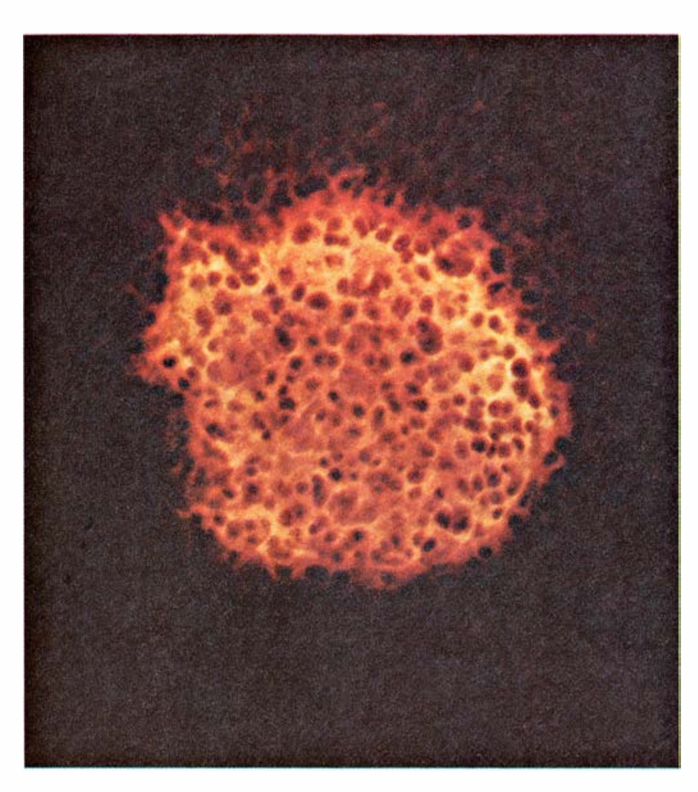
Uno de los principales acontecimientos en la historia de la comprensión de la patología de la diabetes ocurrió en 1889, cuando Oscar Minkowski y el Baron Joseph von Mering, trabajando en Estrasburgo, intentaron ver si el páncreas era esencial para la vida. Mediante un cuidadoso procedimiento quirúrgico consiguieron extirpar totalmente páncreas de perros. Se dice, y posiblemente se trate de una historia apócrifa, que al día siguiente el cuidador del laboratorio se dio cuenta de que la orina de los perros atraía a las moscas, que acudieron en tropel. Sea como fuere, Minkowski y von Mering analizaron la orina y encontraron en ella altos niveles de glucosa, lo que indicaba que la extirpación quirúrgica del páncreas provocaba la aparición de un síndrome parecido a la diabetes. Este descubrimiento implicaba que el páncreas secretaba una substancia capaz de regular el metabolismo de la glucosa.

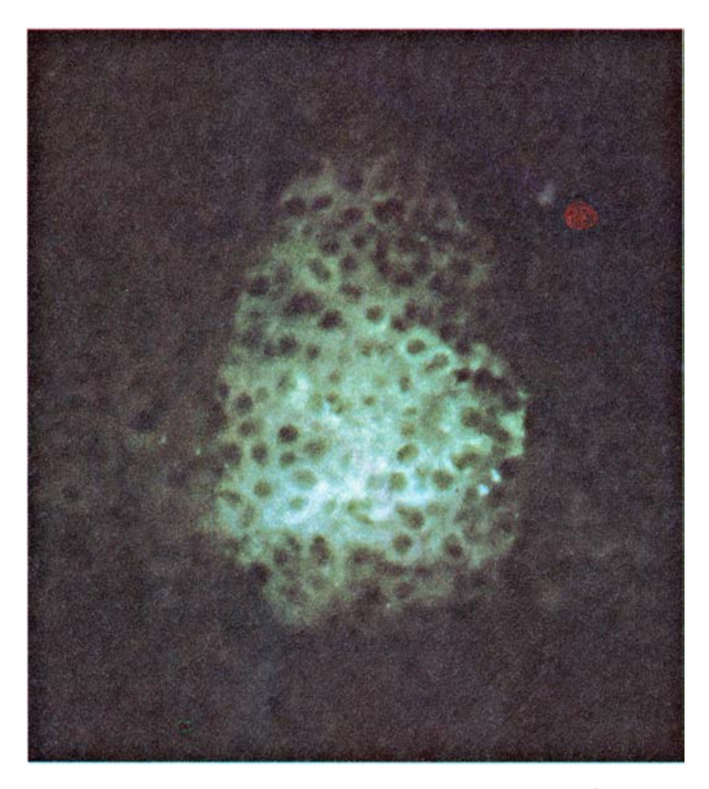
En 1909, la hipotética substancia antidiabética fue bautizada con el nombre de insulina. Todos los esfuerzos por mejorar las diabetes experimentales en perros pancreatectomizados mediante ingestión de páncreas o inyección intravenosa de extractos pancreáticos resultaron, empero, un fracaso. Sabemos ahora que estos experimentos estaban predestinados al fracaso porque la insulina es una proteína que, cuando se administra por vía oral, sufre una degradación enzimática en el tracto digestivo. La misma degradación ocurre en los extractos pancreáticos utilizados para la administración parenteral. La prueba definitiva de la existencia de la insulina no se consiguió hasta que dos investigadores canadienses, Frederick G. Banting y Charles H. Best, extrajeron insulina de páncreas de perros en los que previamente habían eliminado todo tipo de enzimas proteolíticos. El 30 de julio de 1921, Banting y Best inyectaron su extracto pancreático en un perro diabético. A las pocas horas, los niveles de glucosa en sangre empezaron a caer. La noticia del experimento se propagó de inmediato y, en un corto lapso de tiempo, la insulina se empleó de forma generalizada y con gran éxito en el tratamiento de los síntomas agudos de la diabetes mellitus en seres humanos. Se proclamó, entonces, a la insulina como el gran agente capaz de curar la diabetes ya que disminuía los niveles de glucosa en sangre, controlaba los síntomas agudos de la enfermedad y evitaba la muerte por coma que solía sobrevenir a los pocos días del comienzo de los síntomas.

Pero la demostración de que no todo estaba resuelto tardaría algunos años en llegar. Los diabéticos que habían seguido un tratamiento con insulina durante largo tiempo padecían, en una incidencia anormalmente alta, ataques al corazón, muerte súbita, fallos renales, gangrena y ceguera. También eran frecuentes alteraciones nerviosas, de la piel y la boca. Todas estas complicaciones eran especialmente graves durante el embarazo. Así pues, el tratamiento con insulina controlaba los primeros síntomas de la diabetes, pero no la aparición de complicaciones a largo plazo.

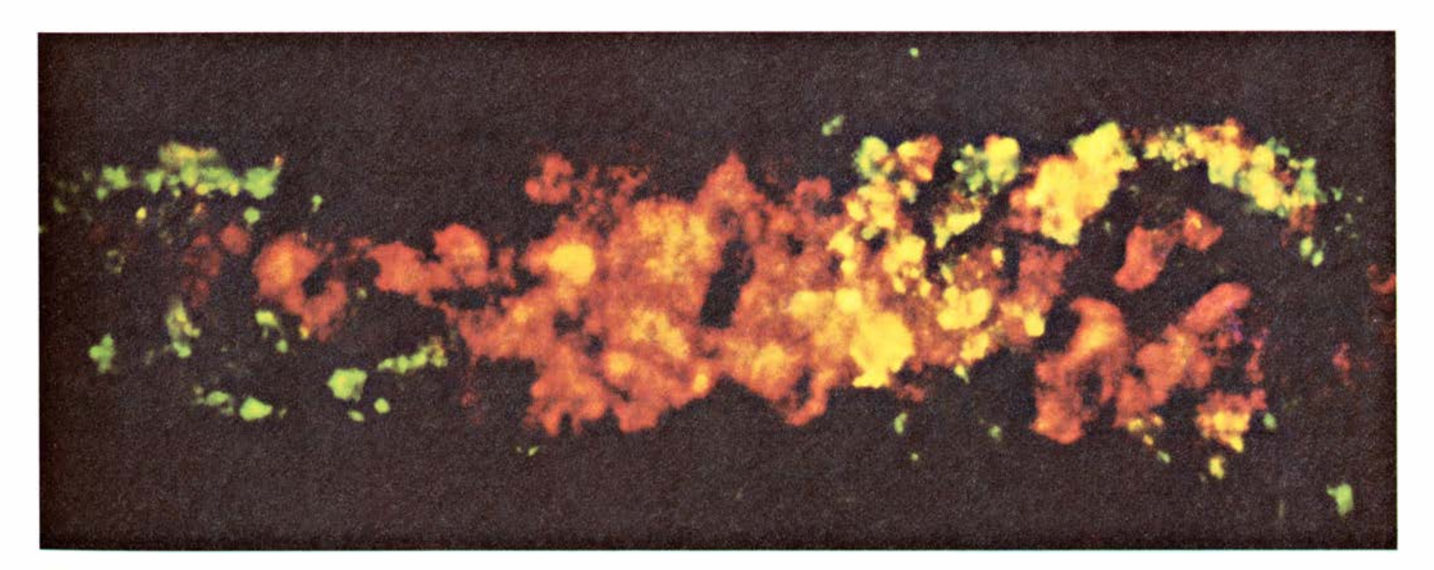
¿Cuáles son las causas de la diabetes y sus complicaciones? Hemos aprendido bastante, desde los tiempos de Banting y Best, sobre la insulina y su modo de regulación de los niveles de glucosa en sangre. El páncreas se compone de dos tipos celulares bien definidos: las células acinares, que producen los enzimas digestivos y los segregan hacia el duodeno (primer segmento del intestino delgado), y los islotes de Langerhans, que segregan diversas hormonas en la corriente sanguínea. En el páncreas se calcula que puede haber entre uno y dos millones de islotes, de aproximadamente 200 micrometros de diámetro y cuya masa, en conjunto, representa un dos por ciento de la masa total del órgano. Los islotes, muy vascularizados, se localizan en la proximidad de capilares.

En cada islote podemos distinguir, al

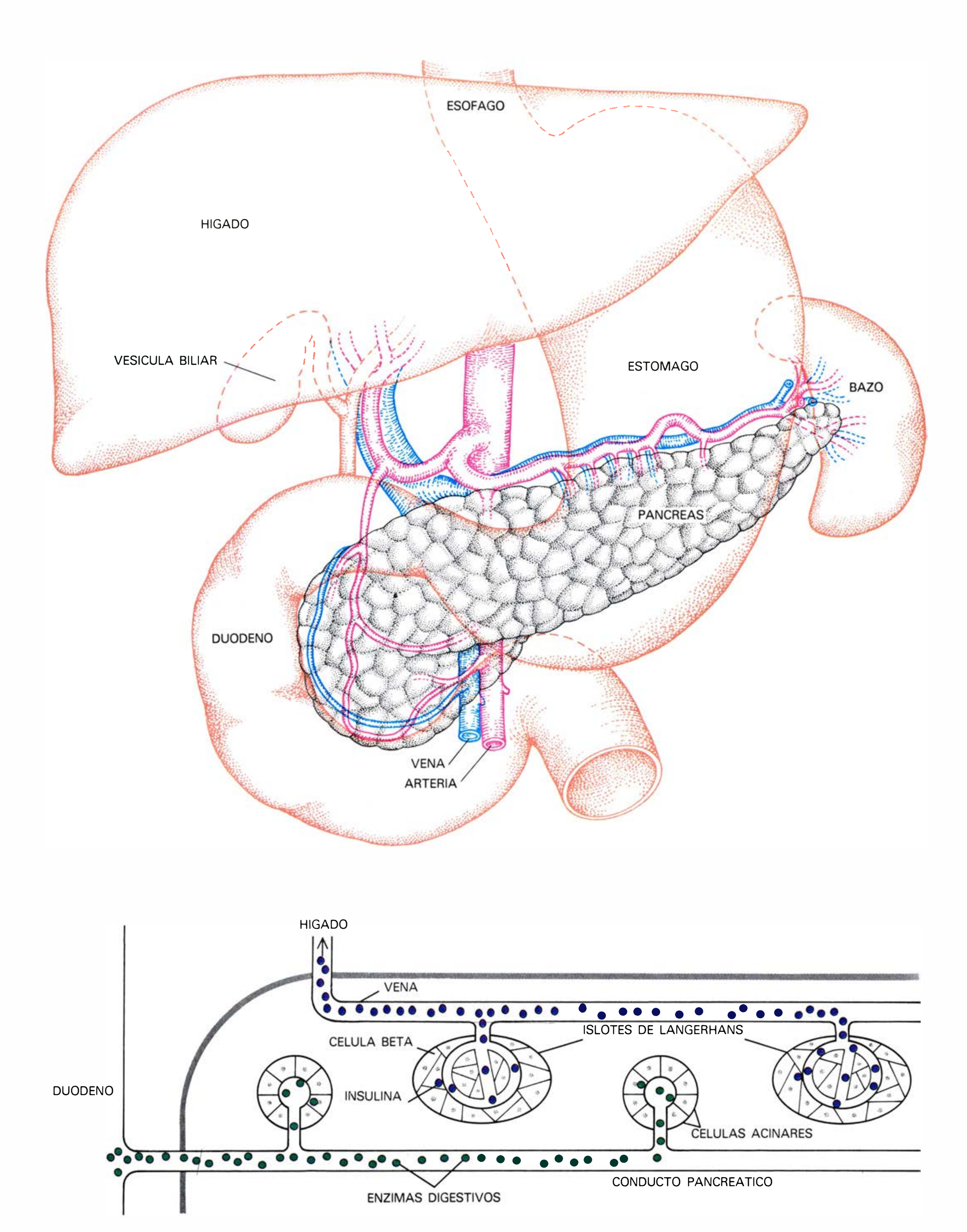




INSULINA contenida en las células beta de islotes de Langerhans del páncreas. Brilla de color naranja en la micrografía de fluorescencia de la izquierda. Las células beta se han hecho visibles tratando una sección de páncreas humano con anticuerpo específico para la insulina y marcado con rhodamina, que colorea de color naranja cuando se ilumina con luz ultravioleta. La zona obscura rodeada por la fluorescencia es el núcleo celular. Las células acinares de alrededor no contienen insulina y, por tanto, aparecen en negro. La micrografía de la derecha es una sección de un páncreas de ratón infectado tres días antes con virus de la encefalomiocarditis (EMC). Se localizó mediante anticuerpos para el virus que tenían unida fluoresceína; éstas reflejan en color verde cuando iluminamos con luz ultravioleta. Sólo las células del islote de Langerhans están infectadas; las células acinares que lo rodean no lo están y aparecen de color obscuro. Estos experimentos han demostrado que los virus pueden destruir la célula beta e inducir diabetes en animales de experimentación. Las micrografías las han tomado A. Bennett Jenson y Kozaburo Hayashi, en el laboratorio del autor.



CELULAS BETA INFECTADAS por virus, reveladas en esta micrografía de inmunofluorescencia. Se trata de una sección de un páncreas de ratón infectado con reovirus y tratado con dos anticuerpos: uno unido a rhodamina, para la insulina, y otro unido a fluoresceína, antivírico. Dependiendo del filtro empleado en el microscopio de fluorescencia, las células beta podían aparecer en naranja y las células infectadas con virus en verde. En doble exposición, se muestra aquí el color mixto de la fluorescencia de suerte que las células beta infectadas por virus aparezcan de color amarillo.



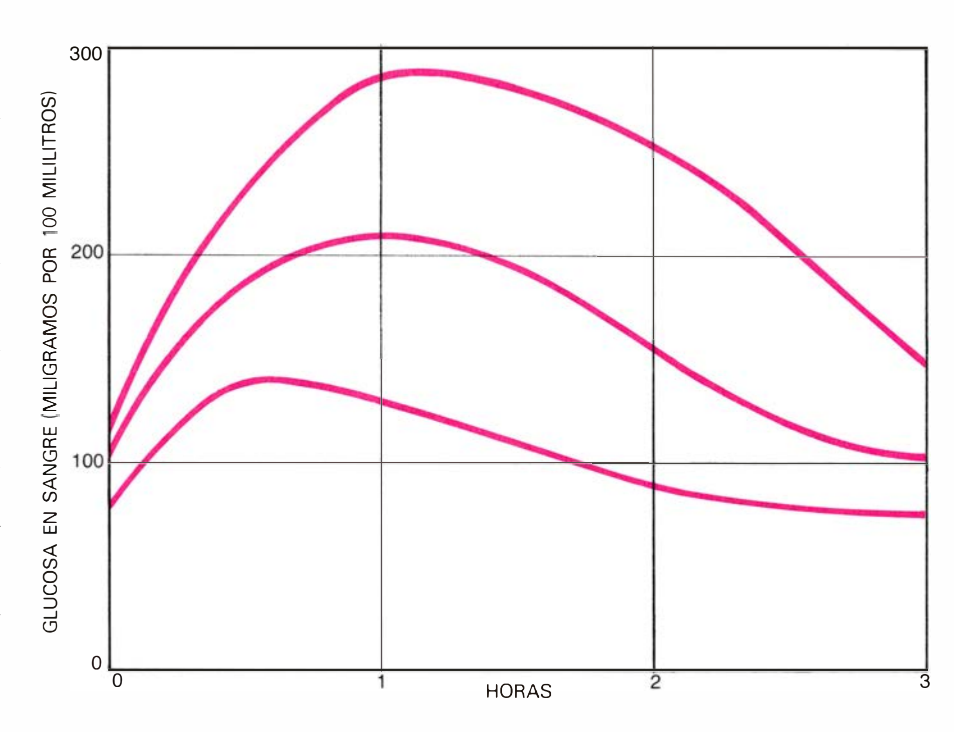
EL PANCREAS está situado en la cavidad abdominal inmediatamente detrás del hígado y debajo del estómago. Está bordeado, a un lado, por el duodeno (primer segmento del intestino delgado) y al otro, por el bazo. El páncreas se compone de dos tipos celulares funcionalmente distintos: las células acinares y los islotes de Langerhans. Las células acinares, que son las más abundantes, fabrican enzimas digestivos que segregan al duodeno a través del conjunto pancreático. Los islotes de Langerhans representan sólo el 1 o 2 por ciento de la masa total del páncreas y segregan varias hormonas (como somatostatina y glucagón) además de la insulina. Todas estas hormonas pasan al torrente sanguíneo a nivel de las pequeñas venas del islote.

menos, cuatro tipos de células diferentes. Las células alfa, que representan el 20 por ciento de la población celular del islote, segregan la hormona glucagón. Las células beta, que forman el 75 por ciento de las células del islote, segregan insulina. La insulina y el glucagón actúan de modo diferente; así, mientras la insulina disminuye los niveles de glucosa en sangre, el glucagón los eleva. Las células delta segregan la hormona somatostatina, que inhibe la secreción tanto de la insulina como la del glucagón, y las células PP segregan la hormona polipeptídica pancreática, cuya función todavía no está muy clara. Aunque la concentración de glucosa en la sangre se mantiene en unos niveles más o menos constantes gracias a la acción de la insulina, tanto el glucagón como la somatostatina desempeñan también un importante papel regulador.

En individuos normales la concentración plasmática de glucosa suele ser menor de 115 miligramos por 100 milímetros de plasma; pero, en los diabéticos, la concentración es mucho mayor, pudiendo alcanzar en casos muy severos 1000 miligramos por ciento. Precisamente, la dificultad que tienen los diabéticos de asimilar el exceso de glucosa presente en la sangre, después de la ingestión de carbohidratos, permite que se les diagnostique con relativa facilidad mediante un test de tolerancia a la glucosa. El sujeto ingiere por boca una cantidad standard de glucosa, midiéndosele la glucemia periódicamente en las horas siguientes. En un individuo normal, los niveles de glucosa vuelven a los valores basales antes de dos horas. El diabético, sin embargo, alcanza unos niveles de glucosa mucho más altos que el individuo normal, tardando más de dos horas en volver a alcanzar los niveles basales.

Después de la ingestión de una comida y en respuesta a la elevación de la glucemia, la célula beta segrega más insulina al torrente sanguíneo. La insulina viaja hasta diversos órganos del cuerpo, interaccionando con receptores específicos que se encuentran en la superficie de las células diana. La unión de la insulina a sus receptores desencadena una serie de hechos que tienen como resultado final el paso de la glucosa al interior de la célula, donde se utiliza para obtener energía o se almacena en forma de glucógeno (almidón animal) y grasa.

Es obvio que cualquier proceso patológico que intervenga en uno o más de los pasos mencionados (páncreas, sangre y tejidos periféricos) puede provocar una alteración en el metabolismo de la glucosa. Por ejemplo, la glucosa en sangre se elevará si la célula beta no fuera capaz de segregar la insulina necesaria, o si hubiera en la sangre antagonistas de la insulina que impidieran su acción, o si los tejidos periféricos no respondieran en la forma adecuada a la acción de la insulina.



TEST DE TOLERANCIA A LA GLUCOSA, método habitual para diagnosticar la diabetes en individuos cuya concentración de glucosa no se eleva de un modo inequívoco. Se administran 75 gramos de glucosa por vía oral y se controlan los cambios en la glucemia periódicamente durante varias horas. La curva inferior muestra la respuesta de un individuo normal. La curva del centro constituye la respuesta de un individuo que sufre alguna alteración en la asimilación de la glucosa, pero que no se le considera verdadero diabético. La curva superior representa la respuesta de un diabético en la que los niveles de glucosa en sangre (glucemia) permanecen elevados a las dos horas de la administración de la dosis.

Las causas de las complicaciones a largo plazo de la diabetes son aún más desconcertantes. Una de las muchas complicaciones consiste en el engrosamiento de la membrana basal que rodea la pared de los capilares. Este engrosamiento contribuye a dificultar la circulación periférica, siendo, al menos parcialmente, responsable de que muchos diabéticos sufran al mismo tiempo alteraciones en más de un órgano del cuerpo. Hay un gran número de hipótesis que intentan explicar la aparición de las complicaciones. Una de ellas sostiene que, en la diabetes, se produce un envejecimiento celular prematuro, de tal forma que la disminución de la función que cabría esperar se produjera en estadios más avanzados de la vida aparece mucho antes. Asimismo, los cambios característicos del envejecimiento se han observado cuando células de diabéticos se han cultivado en el laboratorio, indicando que tales cambios podrían ser controlados genéticamente. Otra hipótesis sugiere que metabolitos específicos intermediarios de la glucosa, caso del sorbitol, se acumulan a altas concentraciones en tejidos tales como nervios y el cristalino del ojo. Este acúmulo podría conducir a cambios en la presión osmótica, provocando un aumento de volumen celular con el consiguiente daño en el tejido.

Ultimamente se está prestando mucha atención a la hipótesis que sostiene que las altas concentraciones de glucosa en sangre permiten a las moléculas de glucosa formar enlaces químicos con los grupos amino de las proteínas celulares, en la reacción conocida como glucosilación. Este hecho se comprobó cuando se encontró que en la sangre de los diabéticos había, con una frecuencia anormalmente alta, formas glucosiladas de hemoglobina. Muchos investigadores están comenzando ahora a pensar que las proteínas de otros tejidos, tales como los nervios, el ojo y los vasos sanguíneos, pueden llegar a glucosilarse en un grado mayor en los diabéticos que en los no diabéticos. El punto crucial es saber si este y otros cambios causados por altas concentraciones de glucosa en sangre son los auténticos responsables del engrosamiento de la membrana basal de la pared capilar y de las otras complicaciones a largo plazo de la diabetes. La cuestión no es ningún bizantinismo académico; si se pudiera demostrar una re-

lación de causalidad entre los altos niveles de glucosa en sangre y las complicaciones de la diabetes, la medida más prudente a adoptar durante el tratamiento sería el control riguroso de la glucemia para mantenerla siempre en un nivel lo más cercano posible a la normalidad.

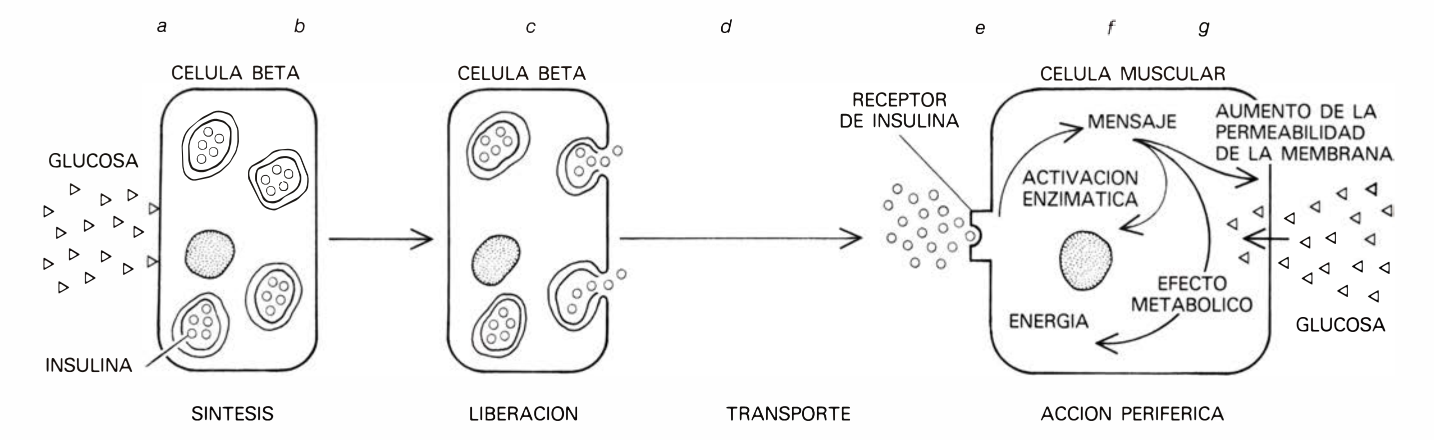
Los diferentes y a menudo contrapuestos descubrimientos sobre la naturaleza de la diabetes y sus complicaciones a largo plazo han llevado a creer que no es una enfermedad única sino, más bien, un grupo heterogéneo de enfermedades, las cuales, en última instancia, conducen a una elevación de la glucemia. De acuerdo con las pruebas clínicas, pueden distinguirse dos tipos de diabetes: una forma de comienzo en la madurez y otra de comienzo en la juventud. Las complicaciones a largo plazo se desarrollan en ambos tipos por igual, aunque hay considerables diferencias individuales.

La forma de comienzo en la madurez es la más frecuente. Representa más del 90 por ciento de todos los casos. Casi siempre, el comienzo ocurre en personas de más de 40 años de edad y con sobrepeso. El inicio es lento y no siempre con cambios patológicos aparentes en el páncreas. Además, las manifestaciones clínicas no suelen ser muy espectaculares y los altos niveles de glucosa en la sangre pueden controlarse normalmente mediante la dieta. Aunque la diabetes se ha venido relacionando tradicionalmente con una deficiencia en la insulina, muchas formas de comienzo en la madurez tienen una suficiencia o incluso un exceso de esta hormona en la sangre. En estos casos, el defecto no es de la insulina (de su insuficiencia) sino, probablemente, del sistema de reacciones que la hormona desencadena en las células diana. Así pues, se está de acuerdo en denominar a la forma de comienzo en la madurez como diabetes no dependiente de insulina (abreviado a veces no insulin-dependiente).

De nuevo aparecen un gran número de hipótesis que intentan explicar la diabetes del adulto, no habiendo unanimidad en la materia. Una de las más interesantes deriva de los estudios llevados a cabo por Jesse Roth, C. Ronald Kahn y sus colegas del National Institute of Arthritis, Metabolism, and Digestive Diseases. Midiendo la unión de la insulina, marcada con un isótopo radiactivo (I^{125}), a sus receptores se encontraba que el número de estos disminuía en pacientes obesos afectos de diabetes del adulto. Sin embargo, cuando el paciente se sometía a una dieta hipocalórica para reducir peso, el número de los receptores de insulina volvía a la normalidad. Roth y Kahn dedujeron que el aumento de ingesta de alimentos asociado a obesidad conducía inicialmente a un exceso de la secreción de insulina a la circulación. Esta insulina, mediante algún tipo de mecanismo de contrarregulación negativa, reducía el número de receptores en la célula diana. Esta disminución sería la responsable de la resistencia celular a la insulina haciéndola menos capaz de utilizar la glucosa. Sin embargo, otros investigadores sostienen que el defecto primario se localiza en el mecanismo que activa la acción de la insulina en el receptor, ya dentro de la célula diana. A pesar de estas discrepancias, la mayoría coincide en que el control adecuado de la dieta y del peso corporal son dos factores de gran relieve en el tratamiento de este tipo de diabetes.

La diabetes de comienzo juvenil es mucho menos frecuente que la del adulto. Representa menos del 10 por ciento de todos los casos. Suele desarrollarse en personas menores de 20 años y tiene un inicio más brusco. La enfermedad se caracteriza por una disminución marcada del número de células beta en el páncreas (por debajo del 10 por ciento de lo normal), que provoca una deficiencia en la producción de insulina y una elevación de la glucemia. El defecto en los niveles de insulina provoca la lipolisis de la reserva grasa del organismo, formándose gran cantidad de cuerpos cetónicos y ácidos grasos. Estos metabolitos disminuyen el *p*H de la sangre, produciendo el fenómeno conocido como cetoacidosis diabética, que puede conducir a la muerte. Debido a que son necesarias inyecciones periódicas de insulina para regular la glucemia, a esta forma de diabetes, la cual es generalmente más severa, se le denomina diabetes dependiente de insulina (que se abrevia insulin-dependiente).

Aunque no hay acuerdo general acerca de las causas de esta forma juvenil, la reducción del número de células beta y la disminución de los niveles plasmáticos de insulina sugerían, desde hace tiempo, que el defecto primario se encontraba en la célula beta. En los últimos años se ha obtenido nueva información sobre los factores que podrían ser



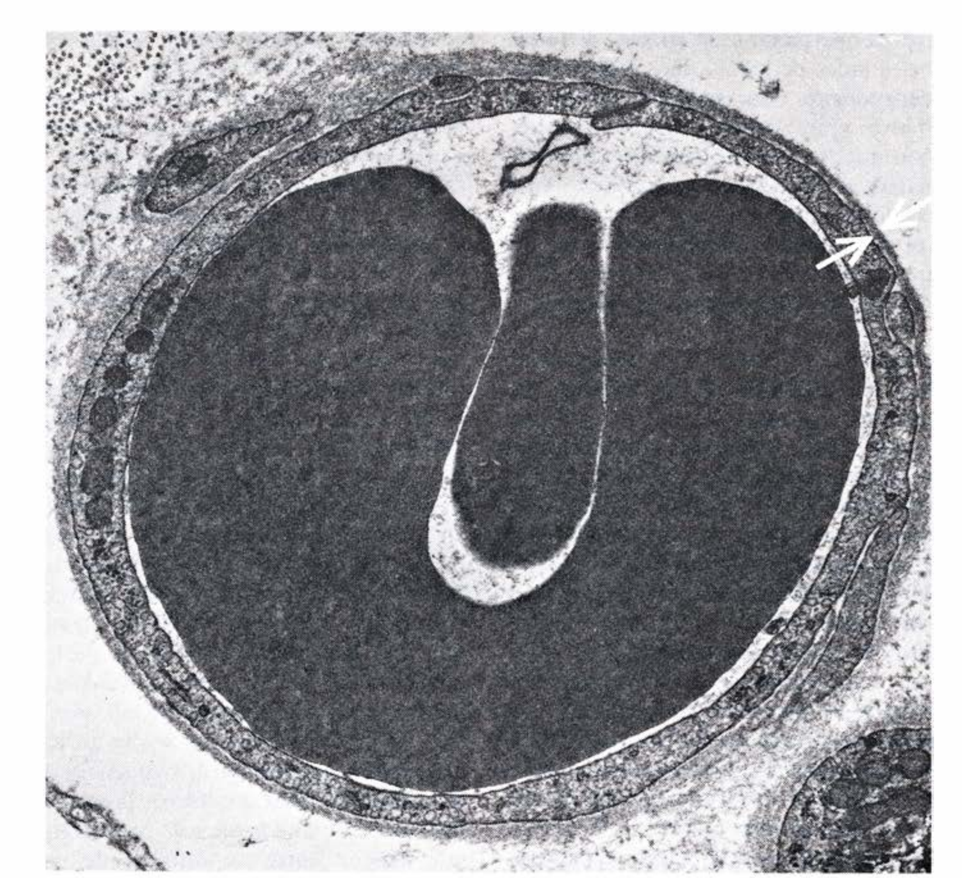
ACCION DE LA INSULINA, esquematizada. La elevación de los niveles de la glucosa en sangre después de una ingesta de carbohidratos induce que las células beta de los islotes de Langerhans segreguen insulina a la circulación. La insulina viaja a lo largo del torrente sanguíneo hasta las células diana en las que actúa, uniéndose a moléculas receptoras de la superficie celular. Esta interacción desencadena una serie de acontecimientos en el interior de la célula que facilitan la entrada de glucosa procedente de la circulación y su posterior metabolismo para obtener energía o su almacenamiento en forma de glucógeno (almidón animal) y grasa. Cualquier defecto en esta cadena de acontecimientos podría producir diabetes. Las posibles causas van desde la destrucción de las células beta (*a*), hasta la síntesis anormal de la hormona insulina (*b*), pasando por la liberación retardada de la insulina (*c*), inactivación de la insulina en la circulación sanguínea por anticuerpos u otros agentes bloqueadores (*d*), alteración de los receptores de insulina o disminución de su número en las células periféricas (*e*), defecto en la traducción del mensaje posteriormente a la unión al receptor (*f*) y metabolismo anormal de la glucosa (*g*). En la diabetes de comienzo juvenil, el defecto primario se localiza en las células beta de los islotes.

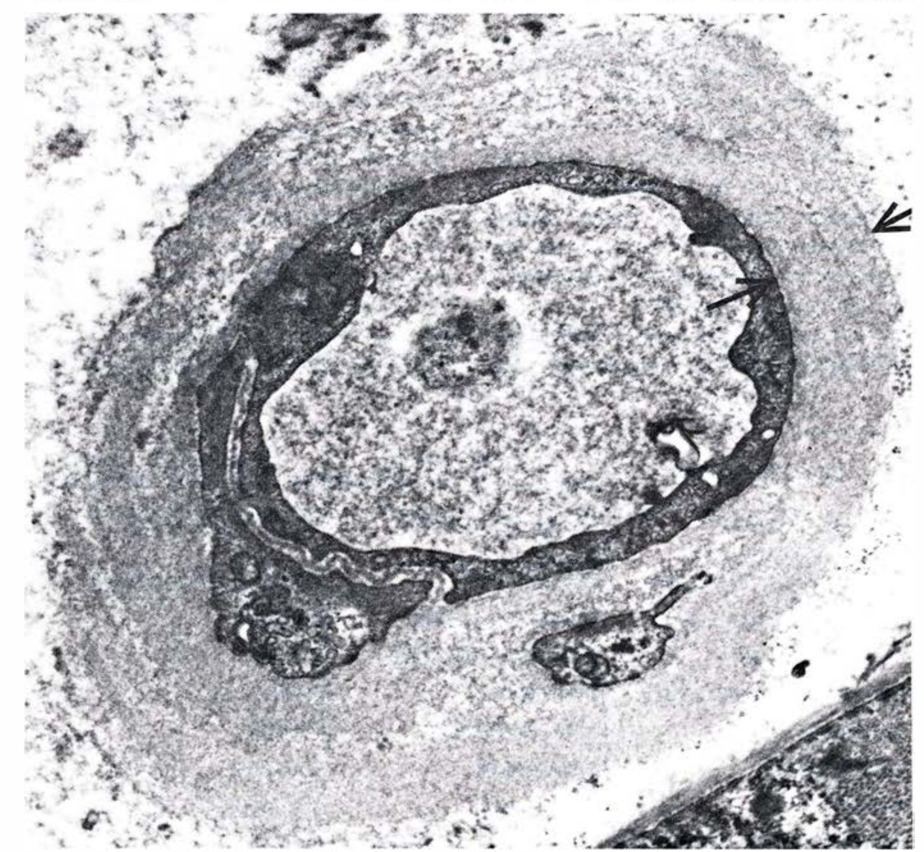
la causa del daño a la célula beta. Estos nuevos aspectos se refieren a la relación entre los factores genéticos y ambientales, sobre los que se discutirá ahora en profundidad.

Desde hace tiempo se sabe que la diabetes del adulto tiene una frecuencia familiar y que la probabilidad de que un individuo desarrolle la enfermedad aumenta si uno de los padres padece el trastorno y, mucho más, cuando son los dos padres. Por el contrario, la probabilidad de padecer la forma juvenil no aumenta en el caso de padres diabéticos. No obstante, los factores de riesgo que influyen en el desarrollo de los diversos tipos de diabetes han sido difíciles de establecer debido a la gran variedad de criterios diagnósticos y a que determinadas variables, como la dieta, edad, sexo, obesidad y rasgos étnicos, no siempre se han tenido en cuenta. Además, el hecho de que en una familia, cuyos miembros están sometidos a las mismas condiciones dietéticas y ambientales, la incidencia de la diabetes sea alta, no prueba necesariamente la implicación de factores genéticos.

En la esperanza de distinguir entre los factores genéticos y ambientales, los genetistas comenzaron, hace ya unos 40 años, a estudiar los gemelos univitelinos: que provienen del mismo óvulo y portan el mismo material genético. Si la diabetes fuese causada exclusivamente por factores hereditarios, al padecer uno de los gemelos la enfermedad, cabría que el otro también la desarrollara. El grado de implicación de los factores genéticos en el desarrollo de la diabetes se puede estimar, por tanto, dependiendo del grado de concordancia (ambos gemelos padecen la enfermedad) frente al grado de discordancia (sólo uno de los gemelos padece la enfermedad).

A comienzos de los años 70, David A. Pyke y sus colegas, del King's College Hospital de Londres, publicaron los resultados obtenidos tras el estudio de más de 100 parejas de gemelos univitelinos (el estudio más amplio realizado con gemelos). Encontraron que, cuando uno de los gemelos desarrollaba la enfermedad después de los 50 años, el otro la desarrollaba también en pocos años, y ello en casi todos los casos. Sin embargo, cuando la diabetes aparecía antes de los 40 años en uno de los gemelos, sólo en el 50 por ciento de los casos se desarrollaba la enfermedad en el otro gemelo. Para más sorpresa, los gemelos que padecían la enfermedad por encima de los 50 años sufrían el tipo de diabetes no-insulin-dependiente, mientras que los





ENGROSAMIENTO de la membrana basal que rodea a los vasos sanguíneos. Se trata de una de los complicaciones a largo plazo de la diabetes. La micrografía electrónica superior muestra un capilar de un individuo normal; la membrana basal (entre las flechas) es una delgada lámina que rodea a las células epiteliales del vaso. (Las estructuras obscuras del interior del capilar son glóbulos rojos.) La micrografía inferior muestra un capilar de un hombre con una diabetes de 26 años de duración. Obsérvese que la membrana basal está notablemente engrosada. Se cree que este engrosamiento contribuye a empobrecer la circulación periférica, dañar la retina del ojo y acelerar el proceso de aterosclerosis. Las micrografías, de 23.000 aumentos, fueron tomadas por Joseph R. Williamson, de Washington.

que comenzaban el desarrollo de la diabetes antes de los 40 años eran insulindependientes. Los resultados de Pyke armaron gran revuelo, ya que demostraban que los factores genéticos predominaban en la diabetes de comienzo en la madurez, mientras que otros factores (probablemente ambientales) eran los que desencadenaban la aparición de la diabetes de comienzo en la juventud.

Otra vía de acceso a la comprensión del problema de la diabetes, distinta de la genética, ha sido investigar la relación entre la diabetes y los antígenos de histocompatibilidad. Estos antígenos (proteínas de la superficie de todas las células nucleadas del organismo) son responsables de que un tejido trasplantado de un individuo a otro sea reconocido como extraño y rechazado por el sistema inmune del receptor. En el hombre, a los antígenos de histocompatibilidad se les denomina sistema HLA. Los genes que codifican los antígenos HLA se encuentran en el cromosoma número 6 y ocupan 4 *loci* a lo largo del cromosoma, que se denominan *A, B, C* y *D*. Los genes de un locus determinado no son siempre los mismos; se denominan alelos las diferentes formas de cada gen. Un individuo puede tener dos alelos diferentes para un locus dado, un alelo aportado por cada progenitor. En el sistema HLA, ambos alelos se expresan como proteínas de la superficie de la membrana celular, las cuales pueden identificarse mediante tests en el laboratorio y servir de base para la tipificación de tejidos para trasplantes.

A los pocos años de que la tipificación de tejidos llegase a ser una técnica usual, se hizo un descubrimiento totalmente inesperado. Ciertos antígenos HLA se encontraban con una frecuencia inusitadamente alta en pacientes con enfermedades específicas. Así, por ejemplo, el riesgo de padecer una enfermedad deformante de la columna vertebral, como es la espondilitis anquilopoyética, era de 100 veces mayor en enfermos portadores del antígeno *B* 27, perteneciente al sistema HLA. Esta observación hizo que gran parte de los investigadores del mundo se pusieran a buscar una correlación entre antígenos HLA y diversas enfermedades, entre ellas la diabetes. Jørn Nerup y sus colegas, del Steno Memorial Hospital de Copenhague, encontraron que los antígenos *B* 8 y *B* 15 eran de dos a tres veces más frecuentes en diabéticos que en no diabéticos. Profundizando en este hecho se dieron cuenta que este aumento de frecuencia de los antígenos *B* 8 y *B* 15 estaba asociado exclusivamente a la diabetes juvenil, mientras que en la diabetes del adulto no había ningún cambio en la frecuencia de los antígenos HLA asociados. Pronto se hizo evidente que había una correlación aún mayor entre la diabetes de comienzo juvenil y los antígenos HLA del locus *D*. Y no sólo eso. Cuando más de un alelo de alto riesgo estaba presente en un mismo individuo, la probabilidad de desarrollar una diabetes juvenil se incrementaba hasta unas 10 veces.

Más pruebas que demostraban la asociación entre la diabetes juvenil y el sistema HLA las aportaron A.G. Cudworth, del St. Bartholomew's Hospital de Londres, José Barbosa, de la Facultad de Medicina de la Universidad de Minnesota, y Pablo Rubinstein, del New York Blood Center. Todos ellos estudiaron familias en las cuales había dos o más hermanos que padeciesen diabetes juvenil. Al comparar hermanos diabéticos con no diabéticos, dentro de la misma familia, descubrieron que aquellos habían heredado grupos idénticos de alelos del sistema HLA en un porcentaje mucho mayor que los hermanos no diabéticos. Teniendo en cuenta que se han identificado diferentes alelos de alto riesgo, se puede afirmar que uno o más genes en estrecha proximidad al complejo HLA situado a lo largo del cromosoma número 6 pueden ser determinantes importantes de la aparición de la diabetes juvenil. No obstante, la situación se ha complicado recientemente al encontrarse que determinados alelos del sistema HLA se asocian a una disminución en la incidencia de diabetes juvenil; ello implica la existencia, al mismo tiempo, de genes protectores.

El mecanismo por el cual dichos genes podrían actuar, bien causando la diabetes o previniéndola, todavía no está claro. Hugh O. McDevitt, de la Facultad de Medicina de la Universidad de Stanford, y Baruj Benacerraf, de la Facultad de Medicina de Harvard, han aportado algunos datos de interés. Observaron que determinados genes específicos del ratón controlaban la respuesta inmune (Ir), y que esos genes se localizaban en la misma región del cromosoma que los genes de la mayoría de los antígenos de histocompatibilidad. Sobre esta base es concebible que alteraciones en la respuesta inmune controladas genéticamente puedan influir en la aparición de la diabetes. Por ejemplo, los alelos de alto riesgo asociados al sistema HLA podrían actuar de forma que la respuesta inmune fuera deficiente frente a agentes que atacan preferentemente a la célula beta, de forma que permita que ésta quede dañada y se produzca la diabetes correspondiente. Por el mismo motivo, los alelos protectores podrían mejorar la respuesta inmune frente a tales agentes invasores.

Existen otros caminos por los que las alteraciones en la respuesta inmune podrían favorecer la aparición de la diabetes. Así, bajo ciertas circunstancias, los anticuerpos atacan a sus propias proteínas del organismo produciendo daño tisular; este fenómeno se conoce como autoinmunidad. Hace cinco años, G. Franco Bottazzo, del Meddlessex Hospital Medical School de Londres, Richard Lendrum, del St. Mary's Hospital de Londres, y James C. Irvine, del Royal Infirmary de Edimburgo, descubrieron

CARACTERISTICAS	DIABETES JUVENIL (DEPENDIENTE DE INSULINA)	DIABETES DEL ADULTO (NO DEPENDIENTE DE INSULINA)
EDAD DE COMIENZO	MENOS DE 20 AÑOS	SOBRE LOS 40 AÑOS
PROPORCION DE LA POBLACION DE DIABETICOS	MENOS DEL 10 POR CIENTO	MAS DEL 90 POR CIENTO
INCIDENCIA ESTACIONAL	OTOÑO E INVIERNO	NINGUNA
SINTOMATOLOGIA	AGUDA O SUBAGUDA	SUAVE
CETOACIDOSIS METABOLICA	FRECUENTE	RARA
OBESIDAD	POCO FRECUENTE	COMUN
CELULAS BETA	DISMINUIDAS	VARIABLE
INSULINA	DISMINUIDA	VARIABLE
INFLAMACION DE LOS ISLOTES	PRESENTE AL COMIENZO	AUSENTE
ANTECEDENTES FAMILIARES	POCO FRECUENTES	COMUN
ASOCIACION A HLA	SI	NO
ANTICUERPO ANTI-ISLOTE	SI	NO

DIFERENCIAS entre la diabetes de comienzo juvenil y la diabetes de comienzo en la madurez. Esta última, que aparece con mayor frecuencia, es menos severa. No requiere inyecciones de insulina.

GENES DEL COMPLEJO HLA	*D*	*B*	*C*	*A*
ALELOS ASOCIADOS CON AUMENTO DE SUCEPTIBILIDAD	*DW* 3 *DW* 4 *DRW* 3 *DRW* 4	*B* 8 *B* 15 *B* 18 *B* 40 *BW* 22	*CW* 3	*A* 1 *A* 2
ALELOS ASOCIADOS CON DISMINUCION DE SUCEPTIBILIDAD	*DW* 2 *DRW* 2	*B* 5 *B* 7		*A* 11

ANTIGENOS DE HISTOCOMPATIBILIDAD: se trata de proteínas de la superficie celular que proveen a todos los tejidos de un individuo de un sello propio característico. Estos antígenos, conocidos como sistema HLA, se codifican en una serie de genes del cromosoma número 6 (en el hombre). Los genes de cada posición o locus de dos cromosomas homólogos no son siempre idénticos, sino que hay una gran variedad de formas denominadas alelos. Determinados alelos del sistema HLA parecen hallarse asociados a un aumento o disminución del riesgo de desarrollar diabetes juvenil. Esta correlación implica que genes íntimamente relacionados con los alelos de alto riesgo del sistema HLA (genes que quizá controlen la respuesta inmune) tal vez desempeñan un papel en la génesis de la diabetes.

en el suero de enfermos, recientemente diagnosticados de diabetes juvenil, un anticuerpo que reaccionaba con células alfa, beta y delta de los islotes de Langerhans de individuos sanos, no diabéticos. El porcentaje de enfermos con diabetes juvenil que poseen tales anticuerpos se eleva hasta un 85 por ciento de los casos en el momento del diagnóstico, pero disminuye a menos de un 25 por ciento a los dos años del dictamen. Sin embargo, en los pacientes con diabetes de comienzo en la madurez raramente se detectan anticuerpos anti-islote.

Se especula con la posibilidad de que los alelos HLA de alto riesgo que se asocian con la diabetes de comienzo juvenil influyan en la formación de anticuerpos anti-islote. Pero, ¿qué es lo que desencadena la producción del anticuerpo y qué papel desempeña éste en la formación de la diabetes? Para ciertos investigadores, un desequilibrio en las células del sistema inmune es el responsable de la producción de anticuerpos anti-islote. Otros piensan que el anticuerpo representa una respuesta inmune a componentes del islote alterados por virus o tóxicos químicos. Todavía hay otros que dudan de la importancia de estos anticuerpos en la patogenia de la diabetes, ya que también se encuentran presentes en individuos que no muestran signos específicos de diabetes.

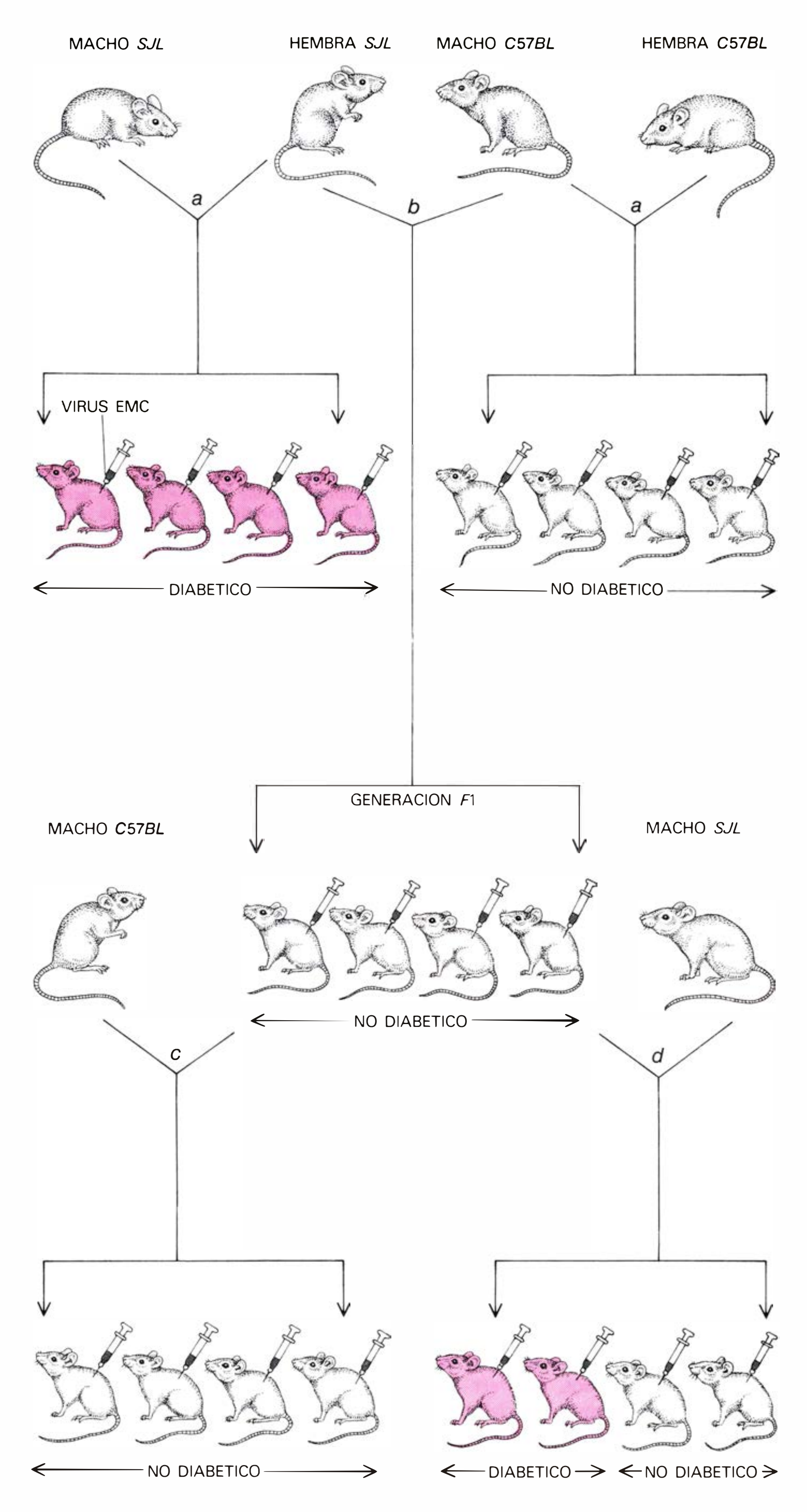
Esta controversia se puede resolver mediante algunos experimentos que se están realizando. Ake Lernmark y sus colegas, de la Universidad de Chicago, han demostrado recientemente que el anticuerpo anti-islote puede reaccionar con antígenos de la superficie de células beta en cultivo. Se intenta desarrollar el sistema de cultivo de las células beta para ver si los anticuerpos pueden destruir realmente a dichas células. Otros investigadores pretenden dilucidar si en el daño celular del islote interviene, no ya el anticuerpo, sino células de la serie blanca sanguínea, del tipo de linfocitos y macrófagos. Independientemente de todas las opiniones vertidas sobre la importancia de los anticuerpos en la patogenia de la diabetes, lo que sí parece claro es que la detección de los anticuerpos anti-islote es un marcador válido para identificar la diabetes juvenil y diferenciarla de la forma adulta.

Tanto los estudios genéticos como inmunológicos antes mencionados aceptan la posibilidad de que un agente ambiental, tipo virus, podría desencadenar el desarrollo de la diabetes juvenil, quizás induciendo una respuesta autoinmune. En 1965, Willy Gepts, de la Universidad de Bruselas, examinó el páncreas de gran número de diabéticos (forma juvenil) que habían muerto poco después de la aparición de los primeros síntomas de la enfermedad. Gepts encontró en los islotes de Langerhans de muchos de estos pacientes células de la serie blanca que podrían hacer pensar en la existencia de una infección o de una reacción autoinmune. Otros datos que pueden apoyar esta posibilidad de proceso infeccioso son el comienzo, frecuentemente agudo, de la forma juvenil, así como el aumento de su incidencia en otoño e invierno, cuando las enfermedades infecciosas son más comunes.

Fue H. F. Haris el primero que relacionó, a mediados de siglo en Filadelfia, la diabetes con un virus, al observar cómo un paciente desarrollaba una diabetes poco después de padecer paperas. A partir de entonces, han aparecido informes de tanto en tanto en los que se relaciona la aparición de diabetes tras determinadas infecciones víricas. La infección habitual suele ser las paperas (parotiditis), pero no se dispone de una demostración firme que asegure que dicha relación no es fruto de la casualidad. Si el virus de la parotiditis es capaz de infectar a las células beta y producir diabetes en seres humanos, debe ser bajo unas circunstancias muy especiales. Tal vez se trata de una forma poco común del virus, o que el organismo esté genéticamente determinado a ser susceptible a dicho virus.

En un reciente trabajo de Margaret Menser, Jill Forrest y sus colegas de Australia informaron acerca de la posibilidad de que el virus de la rubeola concurra también como causa de diabetes. Observaron que los niños y adultos que habían contraído la rubeola mientras se encontraban todavía en el útero materno, tenían una incidencia de diabetes considerablemente superior a la normal. No obstante, debido a que la rubeola produce la aparición de un gran número de malformaciones congénitas, las causas de la diabetes en estos individuos pueden ser muy diversas y, todavía, no claras.

Probablemente, los mejores estudios que relacionan la diabetes con un agente infeccioso se han realizado en animales de experimentación. En 1968, John E. Craighead, que está ahora en la Facultad de Medicina de la Universidad de Vermont, trabajó con una variante del virus de la encefalomiocarditis (EMC). Este virus causa una encefalitis y una miocarditis en el ratón, y una enfermedad febril en seres humanos. Craighead encontró que muchos ratones infectados con el virus desarrollaban una diabetes. Cuando examinó el páncreas de los animales infectados halló signos inflamatorios en los islotes de Langerhans y, dañadas, muchas de las células beta. Más pruebas en este sentido las aportó Kozaburo Hayashi, compañero mío del National Institutes of Health. Hayashi preparó un anticuerpo para el virus EMC y lo unió a fluoresceína, la cual emite una fluorescencia de color verde brillante cuando se irradia con luz ultravioleta. Extrajo el páncreas de los animales infectados e incubó fracciones del órgano en presencia del anticuerpo marcado. El anticuerpo se unió sólo a las células que contenían el virus EMC, pudiendo ser identificadas por su fluorescencia. El experimento demostró que el virus EMC



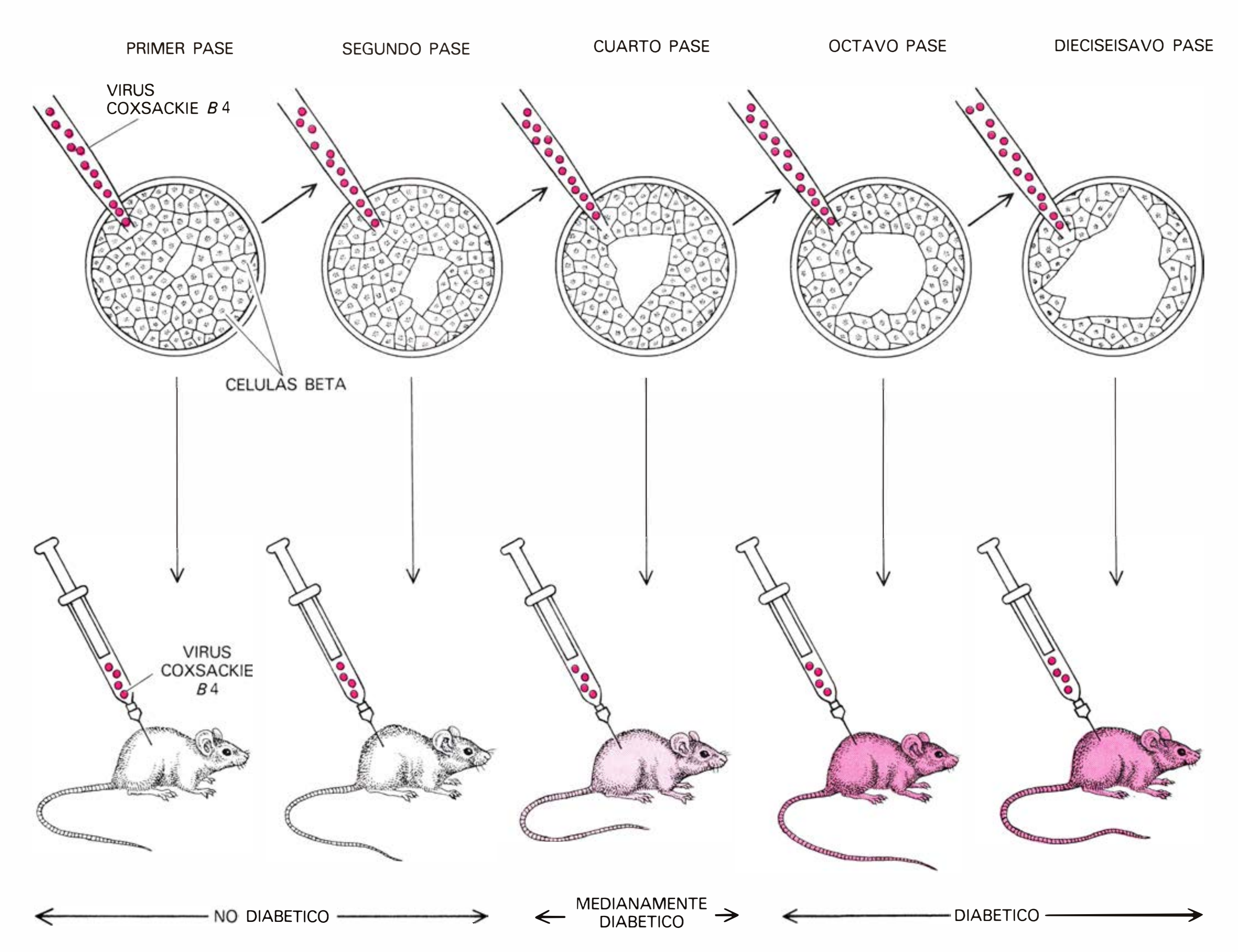
SUSCEPTIBILIDAD GENETICA a la diabetes inducida por virus. Dicha susceptibilidad puede demostrarse en un modelo animal (ratones). Los ratones pertenecientes a la cepa *SJL* desarrollan diabetes cuando se infectan con el virus EMC, mientras que los pertenecientes al tipo *C57BL*, no (*a*). Cuando estos dos tipos –sensible y resistente– se cruzan, su descendencia (la primera generación, F_1) es resistente a la diabetes inducida por EMC (*b*). Igualmente, un cruce retrógrado entre la primera generación (F_1) y el progenitor resistente produce una descendencia resistente (*c*). Pero cuando la F_1 se cruza con el progenitor susceptible, sin embargo, aparece un alto porcentaje de descendencia susceptible (*d*). Ello sugiere que la susceptibilidad a la diabetes inducida por virus EMC se hereda como carácter recesivo.

había infectado las células beta y que la replicación del virus en su interior era la responsable principal de su destrucción.

Trabajos posteriores realizados en el laboratorio de Craighead y en el mío revelaron que el efecto inicial del virus al destruir la célula beta consiste en la liberación de grandes cantidades de insulina a la circulación, disminuyendo los niveles de glucosa en sangre. En pocos días, la reserva insulínica del animal disminuye, en algunos casos a menos de un 10 por ciento de lo normal, y se elevan los niveles de glucosa. En esta fase muchos de los animales empiezan a eliminar la glucosa por la orina y a ingerir cantidades crecientes de agua y comida, los síntomas típicos de la diabetes de comienzo juvenil.

La investigación sobre la implicación de los virus en la génesis de la diabetes tomó un giro distinto cuando se descubrió que no todas las cepas de ratones desarrollaban la diabetes, sino sólo determinados cruces. Tres postgraduados de mi laboratorio, Wark Boucher, Michael Ross y Takashy Onodera, investigaron este fenómeno con detalle. Infectaron a 24 tipos de cruces diferentes de ratón con virus EMC y los dividieron en tres grupos, según fuera la respuesta a la glucosa. El primer grupo, denominado "susceptible", respondía a la infección elevando los niveles de glucosa en sangre. El segundo grupo, "intolerante a la glucosa", sólo elevaba los niveles de glucosa en sangre después de ingerir grandes cantidades de glucosa. El tercer grupo, "resistente", no mostraba ningún signo de diabetes. En los tres casos, siempre después de la infección.

Diversos experimentos de cruzamiento revelaron que las diferencias de susceptibilidad al virus, entre los diversos grupos de ratones, eran controlables genéticamente. La descendencia del cruce de dos ratones "susceptibles" era también susceptible. De forma análoga, la descendencia de dos cepas de ratones "resistentes" era resistente. El cruce entre un ratón del grupo "resistente" y otro "susceptible" producía una primera generación de ratones "resistentes" a la diabetes inducida por el virus EMC. El cruce entre los miembros de esta primera generación daba lugar a una segunda generación entre los que sí aparecían susceptibles, indicando que la susceptibilidad a la diabetes inducida por el virus la habían heredado con un carácter recesivo. En un siguiente experimento se cruzaron los ratones resistentes pertenecientes a la primera generación con otros de la línea del padre resistente (cruzamiento retrógrado), siendo la descendencia resistente: ningún descen-



PASO REPETIDO de un virus a través de células beta en cultivo. Ese pasar constante incrementa su capacidad de producir diabetes. Se cree que, en los pases sucesivos, se van seleccionando las variantes del virus que mejor se reproducen en las células beta. Por ejemplo, el virus Coxsackie *B* 4 normalmente no induce diabetes en el ratón; ahora bien, si se le pasa varias veces su capacidad de destruir células beta y causar diabetes aumenta.

diente desarrolló la diabetes. Pero si los mismos ratones resistentes de la primera generación se cruzaban con la línea del padre susceptible, aproximadamente la mitad de la descendencia era susceptible a la diabetes producida por el virus EMC. Este resultado sugería que era un solo locus génico el responsable de la susceptibilidad.

Como ocurre a menudo en el campo de la ciencia, estos resultados daban respuesta a algunas preguntas, pero motivaban el planteamiento de muchas otras. Ji-Won Yoon, otro científico de mi laboratorio, y Onodera querían saber por qué medio los factores hereditarios influían en la aparición, o no, de la susceptibilidad a la diabetes inducida por el virus EMC. Una posibilidad sería que ciertos genes controlasen dicha susceptibilidad. En los más susceptibles, las células beta estarían infectadas con toda probabilidad y el ratón desarrollaría la diabetes. Para ver si este era el caso, se incubó secciones de páncreas de ratones susceptibles en presencia de anticuerpos para el virus EMC marcados con fluoresceína. El recuento de las células infectadas demostró que los islotes de las cepas susceptibles (tales como las *SJR*) tenían del orden de 10 veces más células beta infectadas que los islotes de los ratones de la cepa resistente (tales como *C5BL*).

Estos experimentos no aclaraban, sin embargo, si la diferencia entre la cepa susceptible de ratones y la resistente se debía a una alteración de la propia célula beta o de la respuesta inmune. A fin de diferenciar ambas alternativas, desarrollamos células beta de cada una de las cepas en unos medios de cultivo apropiados y examinamos su susceptibilidad a la infección. De esta forma pudimos desestimar la respuesta inmune. A todos los cultivos se les infectó con virus EMC, y encontramos que las células beta de las cepas susceptibles eran destruidas más rápidamente que las de las cepas resistentes.

Una indicación acerca de qué es lo que hace que las células beta de determinadas cepas susceptibles sean diferentes, y capaces de ser afectadas por la infección del virus EMC, también lo aportó mi laboratorio. En 1959 John J. Holland, que ahora se halla en la Universidad de California en San Diego, observó que los virus no podían atacar e infectar a ciertas células al menos que poseyeran "receptores" virales: proteínas de la superficie celular que el virus podría reconocer y unirse a ella antes de invadir la célula. Recientemente, Ruben Chairez descubrió, en mi laboratorio, que el virus EMC se unía dos o tres veces más a la célula beta del ratón susceptible que la del resistente. Este descubrimiento sugería que el grado de susceptibilidad al virus podría ser función del número o tipo de los receptores virales de la superficie de la célula beta. Estos experimentos, sin embargo, hay que interpretarlos con cautela: todavía no es posible obtener un cultivo puro de células beta, y el número de receptores virales puede cambiar cuando las células crecen en

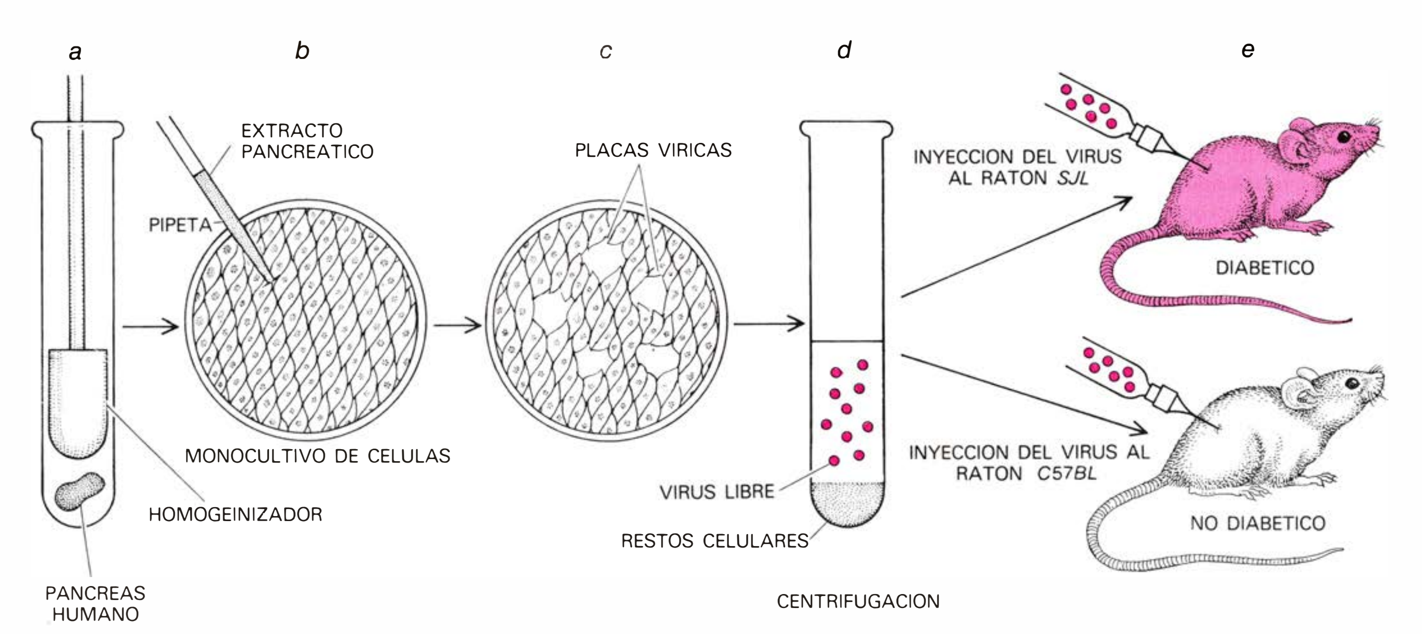
cultivo. Pero ahí está el experimento, en vías de confirmación. Y si ésta se logra, aportará una explicación plausible de por qué tan sólo los animales con la dotación genética "adecuada" desarrollan diabetes inducida por agentes víricos. Es realmente apasionante especular acerca de la posibilidad de que este mecanismo ocurriera de forma similar en el ser humano, y que los alelos de alto riesgo asociados al sistema HLA pudieran controlar la susceptibilidad de las células beta humanas a la infección vírica.

Con la experiencia acumulada del modelo EMC explicado anteriormente, empezamos a buscar otros virus que pudiesen inducir diabetes en el ratón, particularmente aquellos que también infectan a seres humanos. Los primeros candidatos fueron las familias de los Coxsackie. Aislados por primera vez en la década de los 40, a partir de un grupo de pacientes que vivía en Coxsackie, N.Y., esos virus son del tipo ARN (su material genético es el ARN). Pueden producir afección respiratoria en vías aéreas superiores, dolor muscular e invadir el corazón y el cerebro. Poco después de su descubrimiento, Gilbert Dalldorf, del Departamento de Sanidad del Estado de Nueva York, y A. M. Pappenheimer, de la Facultad de Medicina de Harvard, hallaron que los virus del grupo denominado Coxsackie *B* causaban una gran destrucción de los acinos pancreáticos, sin dañar a los inmediatos islotes de Langerhans. Veinte años más tarde, D. Robert Gamble, del West Park Hospital de Epsom en Inglaterra, quien opinaba que quizá hubiera alguna relación entre infecciones por Coxsackie *B* en niños y diabetes, y George E. Burch, de la Facultad de Medicina de la Universidad de Tulane, encontraron destruidos los acinos pancreáticos y ciertos daños en las células beta de los islotes. Gamble observó que había una elevación transitoria de la glucemia en los ratones infectados, pero le resultó muy difícil reproducir los resultados experimentales.

Otros intentos llevados a cabo en mi laboratorio, encaminados a inducir la aparición de la diabetes en ratones con virus Coxsackie *B*, fracasaron también. Entonces, Yoon, Onodera y yo mismo decidimos tomar una senda diferente: hacer el virus más virulento, adaptándolo a la célula beta. Aprovechamos el que las células beta pudieran crecer en cultivo. Infectamos esos cultivos con virus Coxsackie del tipo *B* 4. Después de varios días, tomamos el virus y lo traspasamos a otro cultivo de células beta frescas. Así repetimos el proceso varias veces. Inicialmente el virus se replicaba mal, pero, a medida que realizábamos los pases de un cultivo a otro aumentaba la virulencia, presumiblemente porque en los pases seleccionábamos las variantes del virus que mejor se replicaban en las células beta.

Después de 14 pases con el virus, se lo inyectamos a ratones. En una semana, poco más de la mitad padecía diabetes: tenían disminuidos los niveles de insulina y aumentados los de glucosa en sangre. Muchos de los islotes estaban dañados y, en algunas células beta, se detectaban antígenos virales. Llegamos a la conclusión de que la diabetes que conseguimos inducir se debió no sólo a la adaptación del virus Coxsackie *B* 4 sino, también, a la selección de una cepa determinada de ratones. Como en el caso del virus EMC, sólo ciertas formas del ratón desarrollaban dicha diabetes al ser infectados con el Coxsackie *B* 4. Una vez más, los factores genéticos individuales influían claramente en la susceptibilidad a la diabetes.

Otro virus que estudiamos al objeto de comprobar su capacidad inductora de diabetes fue el reovirus. El prefijo reo (respiratory-entero-orphan) se lo puso Albert B. Sabin (1959), de la Universidad de Cincinnati, porque observó que afectaba al tracto respiratorio y digestivo de niños de los orfanatos. En el ratón ataca las células acinares del páncreas, respetando los islotes. Animados por el éxito de experimento con el Coxsackie *B* 4, lo repetimos con el reovirus. Obtuvimos una cepa de reovirus virulenta para las células beta, que inyectamos a ratones. Cuando examinamos el páncreas de los ratones infectados con reovirus, varios días más tarde, encontramos partículas virales, no sólo en las células beta, sino también en las alfa y en las delta. El número total de islotes dañados por el reovirus era considerablemente menor que los dañados por el Coxsackie *B* 4, y los niveles de glucosa



VIRUS PRODUCTOR DE DIABETES en humanos, aislado, en el laboratorio del autor, a partir del páncreas de un niño que había muerto de diabetes juvenil. Un trozo pequeño del páncreas obtenido en la autopsia se homogeinizó (*a*); se colocó el extracto en una placa de cultivo de células beta que se sabía eran sensibles a una variedad de virus (*b*). Tras varios días, la disolución de las células infectadas por virus dio origen a placas, o agujeros, en la lámina de células; esto último indicaba la presencia de un virus (*c*). Se congelaron y descongelaron varias veces las células para liberar el virus, separándolo mediante el proceso mecánico de centrifugación (*d*). El sobrenadante, que contenía el virus, fue inyectado en ratones susceptibles y ratones resistentes a la diabetes inducida por el virus Coxsackie *B* 4. Días más tarde, los ratones susceptibles desarrollaron una diabetes, pero no así los ratones resistentes (*e*); ello implicaba que la diabetes del niño había sido inducida por el virus que había sido aislado a partir de su páncreas.

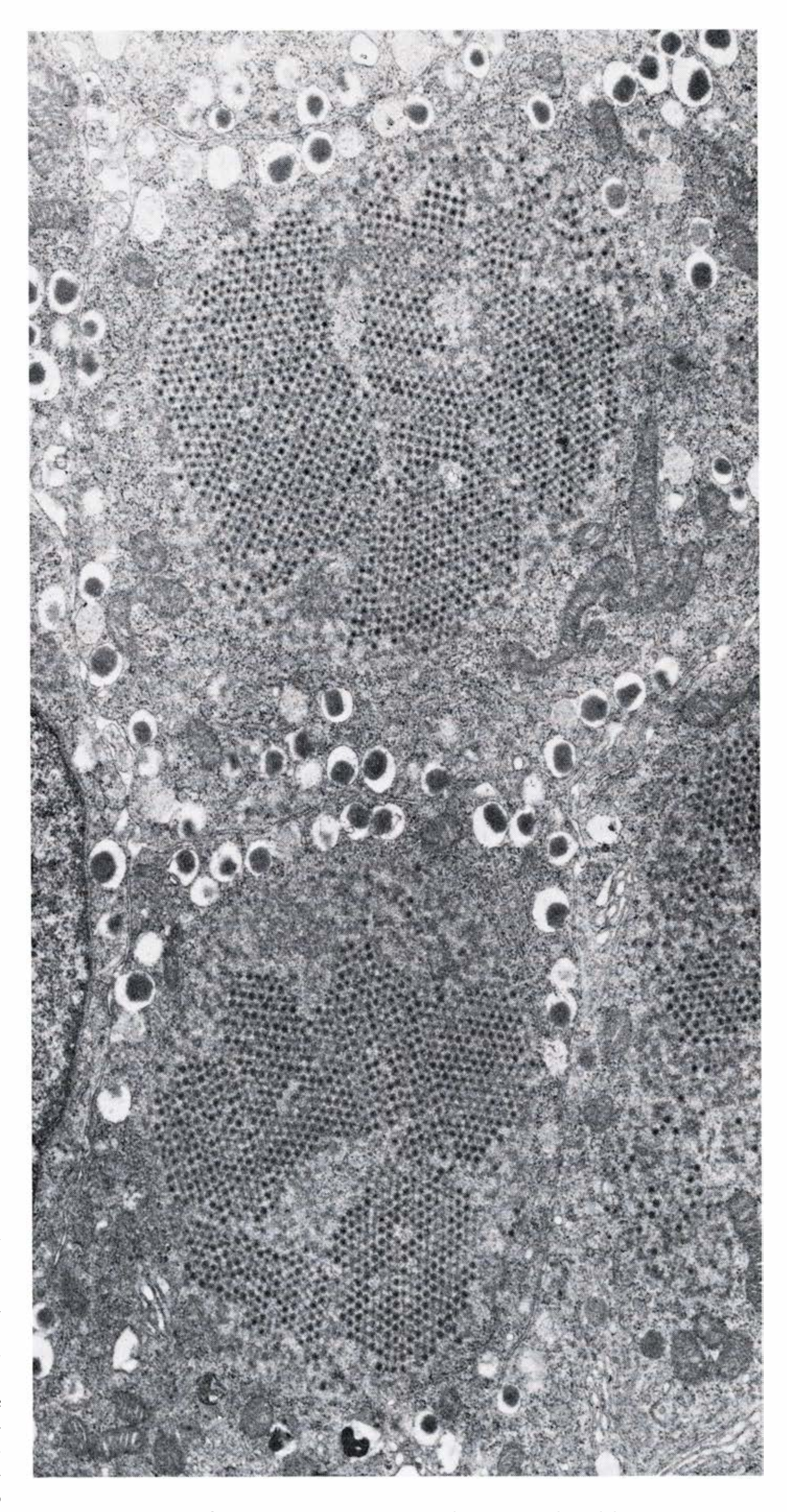
en sangre de la mayoría de los cobayos eran relativamente normales. Cuando sometimos los ratones a una sobrecarga de glucosa sí se hizo patente que había una alteración en su capacidad metabolizadora de la glucosa. Los niveles de glucosa permanecían elevados por un período de tiempo notablemente prolongado, una condición que recordaba la tolerancia anormal a la glucosa por parte de diabéticos humanos.

Elliot J. Rayfield, de la Mont Sinai School of Medicine de Nueva York, observó parecidos sutiles desplazamientos en el metabolismo de la glucosa y la liberación de insulina en hámsters infectados con virus de la encefalitis equina venezolana que, ocasionalmente, produce fiebre, dolor de cabeza y alteraciones del sistema nervioso en seres humanos. Todos estos experimentos, considerados en su conjunto, muestran que variantes de distintos virus comunes que causan patología en humanos pueden atacar las células beta de animales de experimentación e inducir la aparición de síntomas similares a los de la diabetes.

La primera prueba directa de que los virus podían producir diabetes en seres humanos se obtuvo en el año 1978. Un muchacho de 10 años ingresó en un hospital de Washington D.C., con un cuadro gripal de tres días de duración. Había entrado en coma y, al hacerle una analítica, se le descubrió unos niveles de glucosa en sangre elevadísimos (600 miligramos por 100 mililitros); estaba sufriendo una cetoacidosis diabética. A pesar de la intensa terapia a la que se le sometió, su estado general fue deteriorándose poco a poco hasta que falleció. En la autopsia, Robert Marshall Austin, del National Naval Medical Center, descubrió signos inflamatorios en el páncreas y observó que muchas de las células beta habían sido destruidas.

Esos signos concordaban con los observados antes en cobayos infectados con virus que causaban diabetes. Y así, Austin congeló una pequeña muestra del páncreas del niño para su posterior examen. Semanas más tarde, nos contó el caso y nos pusimos de acuerdo para buscar un virus en la pieza. Homogeinizamos la muestra del órgano y añadimos el homogenado a un cultivo de células sensible a una amplia variedad de virus. En pocos días, las células mostraban signos inequívocos de infección. Siguiendo las técnicas usuales en virología, conseguimos aislar el virus. Resultó ser de características similares, aunque no idénticas, a las del Coxsackie *B* 4.

A fin de probar que el virus provenía del paciente y no era una contaminación



PARTICULAS DE REOVIRUS. Forman grandes cristales dentro de las células beta de un ratón infectado, como puede apreciarse en esta micrografía electrónica. También se aprecian vacuolas de secreción, en el citoplasma, que contienen gránulos obscuros de insulina. Aunque la infección por reovirus no produce diabetes severas en el ratón, altera la capacidad del organismo para aprovechar la glucosa del torrente sanguíneo, como puede deducirse de los tests de tolerancia para dicho azúcar.

inadvertida, buscamos pruebas de infección en las muestras de sangre obtenidas del paciente en el curso de su enfermedad. Es bien conocido que, en las primeras fases de una infección viral, la cantidad de anticuerpos antivirales en el suero del enfermo es pequeña o nula, debido a que el sistema inmune del organismo no ha tenido tiempo suficiente de responder a la invasión. Más adelante, sí se pueden detectar fácilmente; un aumento de los niveles de anticuerpo durante un período de varias semanas normalmente implica que la enfermedad está producida por un virus. El suero obtenido del niño diabético cuando fue admitido en el hospital no contenía anticuerpos para el virus coxsackie, pero sí había una cantidad apreciable en el suero obtenido una semana después. Este resultado nos aseguraba que el virus no era una contaminación del laboratorio, pero dejaba abierta la posibilidad de que la infección del niño podría haber sido fortuita y no relacionada con la causa de su diabetes.

En este punto fue donde nuestro modelo de trabajo en animales nos resultó muy útil. A ratones de los que se conocía su susceptibilidad o resistencia a la diabetes producida por el virus Coxsackie *B* 4 inoculamos el virus aislado del páncreas del niño. En una semana, un alto porcentaje de los ratones susceptibles habían desarrollado diabetes, mientras que los resistentes no. El páncreas de los ratones susceptibles tenía signos de reacción inflamatoria, detectándose antígenos virales en las células beta. Estos resultados demostraron que la diabetes del niño había sido inducida por ese mismo virus.

A pesar de todo, la diabetes no parece ser una consecuencia frecuente de la infección por Coxsackie *B* 4. Casi el 50 por ciento de los adultos en los Estados Unidos han estado expuestos una o más veces al virus, sin que la enfermedad haya afectado de un modo claro al páncreas ni se haya producido diabetes. Además, en las formas de diabetes juvenil no se detectan anticuerpos para el Coxsackie *B* 4. El caso que he descrito en las líneas precedentes podría ser el resultado de una rara excepción en la que se combina un virus tóxico para las células beta y una susceptibilidad genética individual. Por otro lado, los virus capaces de atacar a las células beta pudieran ser más frecuentes de lo que se piensa, provocando un alto grado de infecciones subclínicas con cambios mínimos en la célula beta y que, raramente, provocaría un cuadro evidente de diabetes por daño grave en el islote. La situación tal vez sería análoga al efecto del virus de la polio antes de la aparición de las vacunas: muchas personas padecían infecciones subclínicas y sólo unas pocas desarrollaban la forma paralítica. Sin embargo, no hay pruebas de que pueda producirse diabetes por contagio.

Dado que no es frecuente encontrar anticuerpos para el virus Coxsackie *B* 4, en los diabéticos de comienzo infantil, se ha iniciado la búsqueda de otros virus que sí pudieran ser causantes de la enfermedad en el hombre. Cabría que, después de una serie de infecciones virales en la infancia, cada una de las cuales capaz de producir alteraciones en la célula beta, el daño global de los islotes sí causara diabetes, una vez agotadas las células beta. Por otro lado, quizás algunos niños tienen ya, al nacer, un número inferior de células beta o una menor capacidad de reparar o regenerar dichas células una vez dañadas. Si esto fuera así, a los virus les resultaría más fácil producir diabetes en esos niños con deficiencias individuales.

También es posible que los virus fueran una de tantas causas de diabetes (quizá la menos frecuente) y que otros factores ambientales del tipo de tóxicos químicos y drogas tuvieran mayor importancia. En este sentido se sabe desde hace 35 años que la aloxana destruye las células beta y produce diabetes experimental en animales. La droga es muy selectiva en sus efectos: la agresión a la célula beta puede observarse a los pocos minutos de la inyección. Al comienzo de la década de los 60 se vio que otra droga la estreptozotocina, era también tóxica para la célula beta y capaz de inducir una diabetes experimental en animales. Recientemente, Arthur A. Like y Aldo A. Rossini, de la Facultad de Medicina de la Universidad de Massachusetts, han observado que, mientras que, en una dosis fuerte, la estreptozotocina intoxica directamente las células beta de animales de experimentación, múltiples pequeñas dosis parecen actuar de un modo indirecto, de suerte que la célula beta se hace vulnerable al propio sistema inmune del animal. Bajo estas últimas circunstancias, sólo ciertos tipos de ratones mostraban signos inflamatorios en sus islotes y desarrollaban diabetes; ello sugería, una vez más, la importancia de los factores genéticos.

Paradójicamente, a causa de la toxicidad y alta especificidad de la estreptozotocina por la célula beta, se ha utilizado en el tratamiento de unos tumores poco frecuentes de las células beta: el insulinoma. Sin embargo, en general, estos productos tóxicos sólo se utilizan ampliamente en el laboratorio para la producción de diabetes experimentales en animales. En 1975 se introdujo en los Estados Unidos un veneno corrosivo denominado vacor, que tiene una estructura que recuerda a la estreptozotocina. De forma accidental o deliberada, fue ingerido por muchas personas con funestas consecuencias. Algunas murieron y, de los supervivientes, 20 desarrollaron una diabetes aguda que hubo de ser tratada con insulina. En dos autopsias que se realizaron había claramente destrucción de las células beta. Existen otros fármacos capaces de producir toxicidad en la célula beta, aunque el daño es transitorio y no muy grave.

¿Cuántos miles de substancias químicas hay, naturales y sintéticas, a los que el ser humano se encuentra expuesto diariamente y cuyos efectos sobre las células beta no han sido examinados? ¿Algunas de ellas causan daño a la célula beta y producen diabetes de comienzo juvenil? Si eso es así, las causas de la diabetes podrían ser numerosísimas. De cualquier forma no es probable que en un futuro próximo podamos encontrar algo tan sencillo como una vacuna para prevenir la diabetes.

En resumen, las investigaciones realizadas en laboratorios de todo el mundo concuerdan en que la diabetes no es una simple enfermedad con etiología única. La propia diabetes de comienzo juvenil puede tener múltiples causas. De gran importancia es el descubrimiento de que genes cercanos al complejo HLA influyen en el riesgo de desarrollar la diabetes juvenil y que la mayoría de los pacientes diagnosticados últimamente tienen en el suero anticuerpos contra los islotes. La esperanza se cifra en que pueda llegarse a identificar los individuos que son más susceptibles al daño de la célula beta y encontrar la forma de protegerlos.

Muy importante es también el saber que, aunque algunos casos de comienzo juvenil pueden ser explicados por factores genéticos y otros por factores ambientales, hay un tercer tipo en el que es necesaria la interacción entre los factores genéticos, individuales, y los factores ambientales. La relativa importancia de los diferentes insultos ambientales, como virus y tóxicos químicos, los factores genéticos y la autoinmunidad está por dilucidar. Por eso últimamente se está dedicando más atención al tema. Aunque se halla muy lejano el día en que podamos prevenir la aparición de diabetes juvenil y sus complicaciones, los misterios que guarda esta antiquísima enfermedad se vienen desvelando sin solución de continuidad.

Aleaciones con memoria de la forma

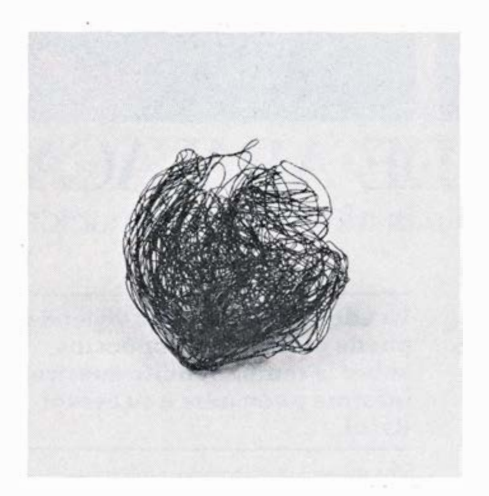
Si a una de estas nuevas aleaciones se le da forma a una temperatura determinada, y se deforma luego a otra temperatura, puede "recordar" su forma original. Este extraño efecto tiene muchas aplicaciones prácticas

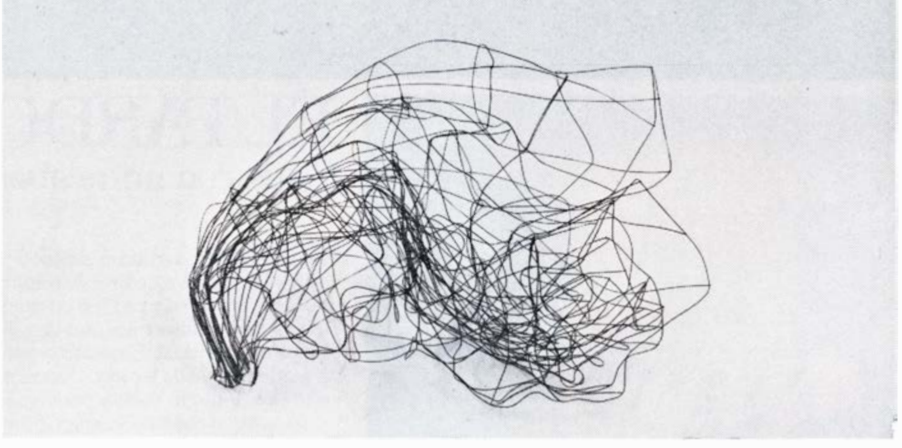
L. McDonald Schetky

Los metales se caracterizan por una serie de propiedades físicas tales como resistencia a la tensión, ductilidad, maleabilidad y conductividad. En el caso de esta nueva familia de aleaciones, se puede añadir ahora a las anteriores propiedades las cualidades antropomórficas de memoria y capacidad de recibir instrucciones. Las nuevas aleaciones presentan el efecto denominado memoria de la forma. Si las aleaciones de este tipo se deforman plásticamente a una temperatura, recuperarán por completo su forma original al someterlas a una temperatura más elevada. Al recuperar su forma, las aleaciones pueden producir un desplazamiento o una fuerza, o una combinación de ambos, en función de la temperatura. Debido a estas notables e insólitas propiedades, las aleaciones con memoria de la forma son útiles para resolver una amplia gama de problemas. En una de estas aplicaciones bien desarrollada por otra parte en la actualidad, estas aleaciones permiten obtener conexiones simples y prácticamente herméticas para conducciones neumáticas e hidráulicas. También se utilizan en los cierres herméticos y en las conexiones de los montajes electrónicos. Su aplicación más reciente ha sido en los sistemas de control mecánicos y electromecánicos, al objeto de que puedan proporcionar, por ejemplo, una respuesta mecánica precisa a cambios de temperatura pequeños y repetidos. Varias aplicaciones prometedoras se están ensayando en medicina. Más remota, pero no por ello menos tentadora, es la posibilidad de aprovechar las aleaciones con memoria de la forma para convertir ciertas fuentes de calor poco importantes (piénsese en el existente en el agua de la superficie oceánica) en energía mecánica.

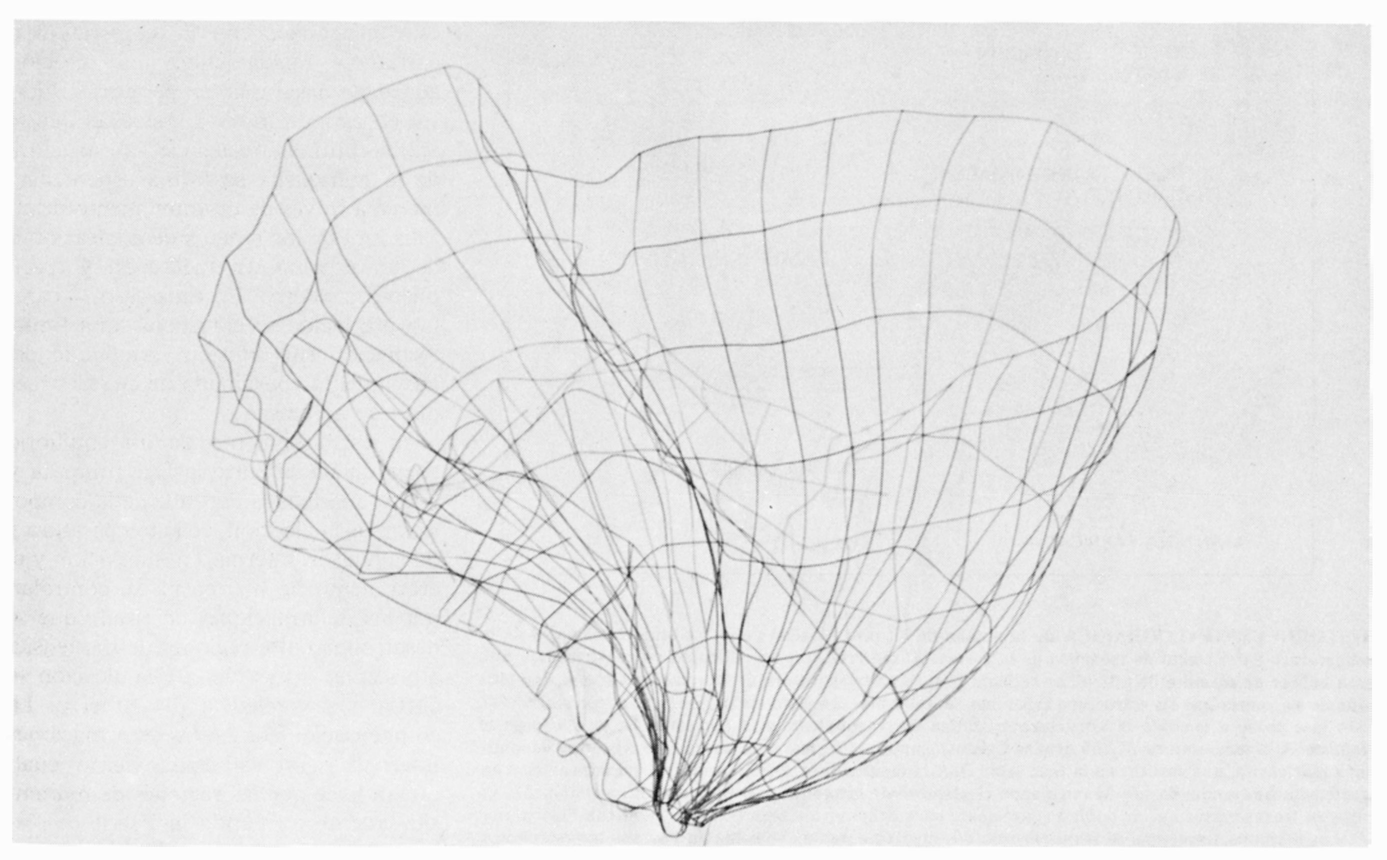
Para que una aleación presente memoria de la forma debe poseer una estructura cristalina capaz de pasar a la configuración conocida con el nombre de martensita al someterla a una temperatura o a un esfuerzo determinados y eliminar luego esa perturbación. Un ejemplo sencillo consistiría en doblar, a temperatura ambiente, un alambre constituido por una aleación con memoria de la forma hasta darle una configuración de trébol de cuatro hojas. Se calienta luego el alambre hasta que su estructura cristalina adquiera una configuración de alta temperatura denominada fase beta o fase primaria ("parent phase"). A continuación, se enfría rápidamente el alambre de modo que los átomos metálicos se reordenen entre sí tomando la forma cristalina correspondiente a la martensita. Ahora se puede doblar o torcer el alambre hasta darle cualquier otra forma. Si más tarde se calienta el alambre hasta una temperatura superior a la correspondiente a la transición de martensita a la fase primaria, se produce un cambio ordenado de grandes grupos de átomos que recompone la forma original de hoja de trébol. Puesto que la transformación en martensita es esencial para que se produzca el efecto de memoria de la forma, las aleaciones que presentan esa propiedad se denominan también aleaciones marmem.

El efecto de memoria de la forma había sido ya observado en 1938 cuando



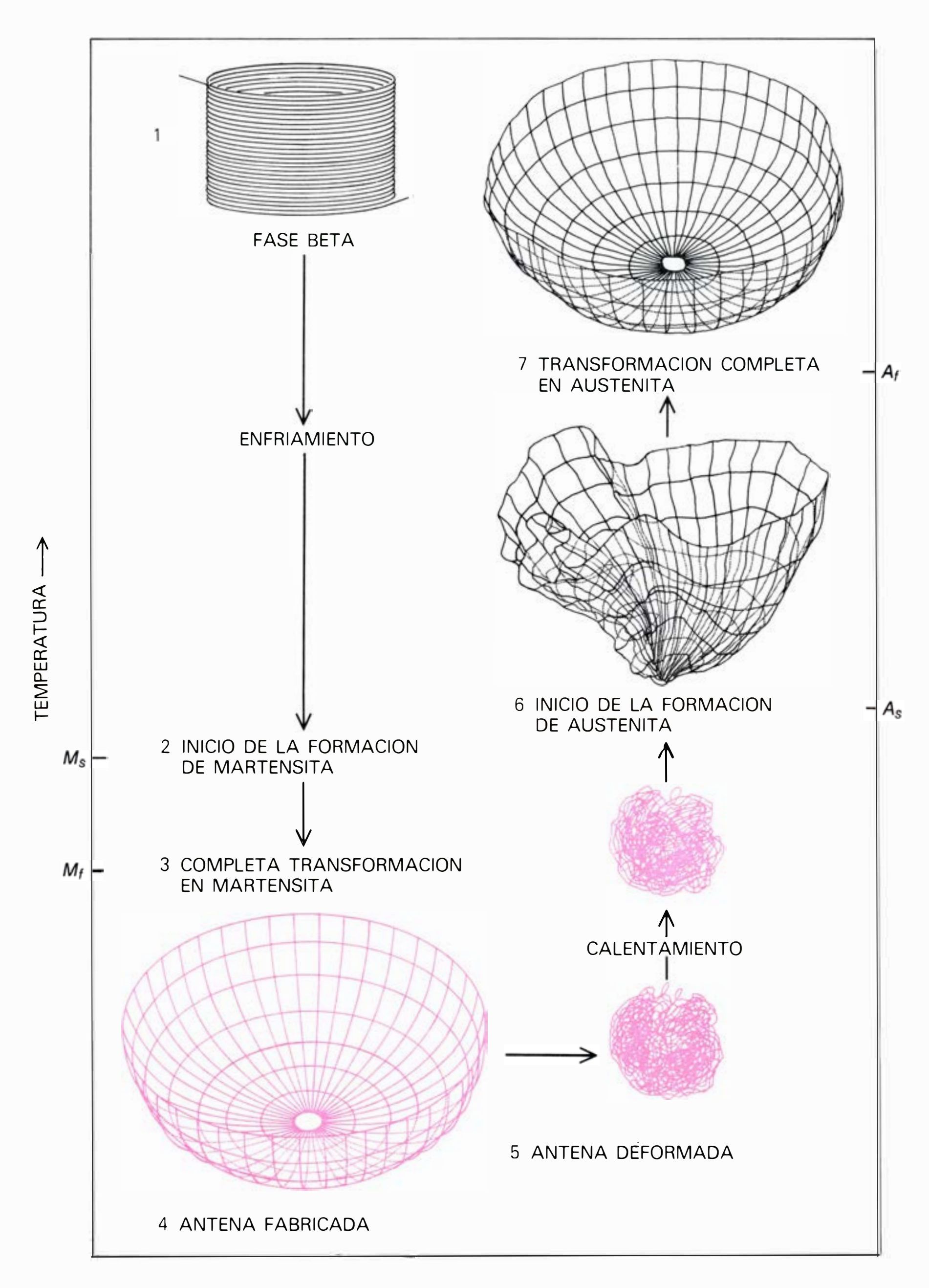


DESPLIEGUE ESPONTANEO de un hemisferio de alambre que podría servir de pequeña antena de un vehículo espacial, en esta secuencia de cuatro fotografías. La antena está hecha de una aleación de níquel y titanio (nitinol), que presenta el efecto de memoria de la forma. En la fotografía de la izquierda, la antena se ha estrujado a temperatura ambiente hasta conseguir una bola apretada de menos de cinco centímetros de diámetro. A medida que va aumentando la temperatura de la masa de alambre, la antena se despliega gradualmente. Cuando la temperatura llega a los 77 grados Cel-



sius, la estructura adquiere su configuración original (*arriba, a la derecha*). El diámetro de la antena desplegada es de unos 25 centímetros. La recuperación de la forma de la antena coincide con la desaparición de cristales de martensita en el material y su sustitución por cristales de austenita, que ofrece una estructura cristalina más simple. La historia cristalográfica de esta secuencia se ilustra en la página siguiente. La antena fue diseñada por la Goodyear Aerospace Corporation. Aunque no ha llegado a entrar en servicio, muestra de una manera clara las propiedades de estas aleaciones.

Alden B. Greninger, de la Universidad de Harvard, y V. G. Mooradian, del Instituto de Tecnología de Massachusetts, demostraron que se podía controlar la formación y la desaparición de la fase martensita del latón (una aleación de cobre y cinc) mediante un cambio de temperatura. Casi al mismo tiempo, G.V. Kurdyumov, metalógrafo ruso conocido por sus trabajos anteriores sobre la cristalografía de las martensitas, especialmente de las del acero, estudió las relaciones de fase en el latón entre la fase beta de alta temperatura y la martensita formada por enfriamiento rápido. Más tarde, Thomas A. Read y sus colaboradores, de la Universidad de Illinois, investigaron el efecto de memoria de la forma en aleaciones de oro-cadmio y mostraron las fuerzas que eran capaces de desarrollar las transiciones de fase. Se ha observado que aleaciones tan distintas como las de hierro-platino, indio-cadmio, hierro-níquel, níquel-aluminio y acero inoxidable presentan la propiedad de memoria de la forma en grado variable.



HISTORIA CRISTALOGRAFICA de la antena de la nave espacial donde se muestra el papel de la temperatura en el efecto de memoria de la forma. El material que se utiliza en su construcción, una gran bobina de alambre de nitinol, se calienta hasta una temperatura de 650 grados Celsius, estabilizándolo de suerte que su estructura cristalina se halla por completo en la fase beta o "primaria" (*1*). Esta fase posee a menudo la estructura cristalina correspondiente a la austenita. Se enfría luego el alambre. A la temperatura M_s (60 grados Celsius) empieza a formarse la nueva fase cristalina denominada martensita, que sustituye a la fase beta (*2*). A la temperatura M_f (52 grados) la transformación en martensita ha terminado (*3*). Manteniendo el alambre de nitinol a una temperatura inferior a M_f, se corta en trozos cortos que se doblan suavemente para originar los segmentos de la antena hemisférica (*4*). Los distintos segmentos se mantienen unidos mediante puntos de soldadura en sus intersecciones. Se puede estrujar entonces la antena hasta reducirla a un pequeño volumen (*5*). Para restituir su forma original, la estructura comprimida se calienta. A la temperatura A_s (71 grados) la austenita empieza a sustituir a la martensita (*6*). Al llegar a A_f (77 grados), la antena se ha desplegado por completo (*7*). En este caso particular, la aleación con memoria de la forma no "recuerda" la verdadera configuración de la antena, sino las curvas suaves del alambre enrollado, utilizado. Cuando el alambre intenta enderezarse, se ve forzado a adquirir la forma de tazón por las múltiples soldaduras de las intersecciones.

No fue hasta 1962 cuando el fenómeno pasó a primer plano. Y ello, gracias al anuncio de la existencia de memoria de la forma en una aleación de niquel y titanio. La aleación fue descubierta por William J. Buelher, del U. S. Naval Ordnance Laboratory (cuyo nombre pasó a ser desde entonces el de Naval Surface Weapons Center) en Silver Spring, Middleessex. La aleación recibió el nombre de nitinol, acrónimo de niquel-titanio y Naval Ordnance Laboratory.

Antes de considerar los distintos modos de utilización del efecto de memoria de la forma describiré con algo más de detalle el fenómeno en sí mismo, las condiciones de temperatura y esfuerzo necesarias para que se produzca y la base cristalográfica que lo rige. La estructura del cristal de martensita, que es esencial para este efecto, puede obtenerse de dos maneras: sometiendo una aleación a un esfuerzo cuya magnitud esté relacionada con la temperatura o enfriando rápidamente una aleación adecuada hasta una temperatura crítica. En el segundo método, que es el que se aplica con más frecuencia, la estructura de la martensita se forma espontáneamente a través de un movimiento de cizalladura de los átomos de la aleación o de un proceso de nucleación y crecimiento cristalino. En cada caso, el cambio producido se considera "una transformación sin difusión" resultante del movimiento coordinado de grandes bloques de átomos.

El establecimiento de un equilibrio termodinámico entre la fase primaria y la fase martensita depende de la composición de la aleación, de la temperatura y de la tensión interna. La nucleación y el crecimiento de martensita se controlan por las deformaciones de cizalla que se desarrollan entre regiones de martensita adyacentes a medida que la aleación se enfría o se somete a un esfuerzo. La compensación entre estas deformaciones internas y un esfuerzo externo cualquiera hace que las regiones de martensita crezcan formando una serie de placas que se adaptan mutuamente. La orientación de una placa en relación con la orientación de su vecina corresponde a la que es energéticamente más estable dentro del campo de deformación parti-

FORMA INICIAL	MOLDEADO ADICIONAL EN FRIO O EN CALIENTE	FORMA TRAS FASE BETA Y ENFRIARSE RAPIDAMENTE	POSICION A TEMPERATURA AMBIENTE	POSICION RECORDADA (POR ENCIMA DE A_f)
	SIN MARTENSITA	AHORA CONTIENE MARTENSITA	MARTENSITA BAJO TENSION	SIN MARTENSITA
	NADA	→ SE DOBLA HASTA CURVARSE		
		→ SE ENDEREZA		

INTERRUPTOR ACCIONADO POR LA TEMPERATURA, diseñado de suerte que se abra, o se cierre, por encima de una temperatura determinada. La temperatura depende de la aleación seleccionada y coincide con la temperatura (A_f) a la que la martensita es sustituida por la austenita. Si se pretende que el interruptor se abra por encima de la temperatura A_f (*serie de dibujos de la fila superior*), se calienta una varilla recta de aleación hasta la temperatura de formación de la fase beta y luego se enfría rápidamente. La varilla contiene ahora martensita (*en color*). La varilla se dobla para que pueda utlizarse como interruptor, con lo que la martensita está sometida a un esfuerzo. Cuando se calienta la varilla por encima de A_f, desaparece la fase martensita y se endereza, cerrando el interruptor. Si se diseña el interruptor para que se abra por encima de A_f, debe doblarse la varilla antes de enfriarla y hacerla pasar a la fase beta (*serie de dibujos de la fila inferior*). En este caso se endereza la varilla antes de colocarla en el interruptor.

cular existente. Tanto si el campo es el resultado de un esfuerzo aplicado, como si obedece a un cambio de temperatura, en el conjunto del material las placas adquieren diversas orientaciones y tamaños igualmente variables.

Si bien la estructura de la fase beta primaria y la de la martensita presentan detalles distintos para cada aleación, la fase típica de temperatura elevada es una fase desordenada con estructura cúbica centrada en el interior (en donde los átomos forman una red cúbica con un átomo en el centro de cada cubo y otros ocho átomos en sus vértices). Cuando se hace bajar la temperatura, esta fase pasa a una estructura que puede estar ordenada, y ser cúbica centrada en el interior, o bien constituir una superred. En la estructura ordenada, los átomos de uno de los componentes de la aleación se sitúan en sitios preferentes de la red cúbica en relación con los átomos de las otras clases. La superred puede visualizarse como celdas interpenetradas, en las que la celda unidad del cristal, es decir, la unidad más pequeña que al repetirse y extenderse en todas direcciones define la estructura cristalina, está constituida por varias docenas de átomos, o más. Esta complejidad de la estructura cristalina hace difícil describir los movimientos relativos producidos en los átomos cuando tiene lugar la transformación en martensita. Se invocan a menudo, para describir estos movimientos, términos como los de cizalla y mezcla interlaminar (como la acción de barajar las cartas). En la aleación tipo marmem constituida por cobre, cinc y aluminio se forman cuatro variantes de martensita, además de la fase beta primaria, ostentando cada una de ellas una orientación cristalina desplazada 60 grados respecto a las otras.

La deformación asociada a una variante compensa la deformación en las otras variantes, y así el crecimiento de campos múltiples mutuamente adaptados de placas de martensita está energéticamente favorecido con respecto al crecimiento de una placa única. Los límites entre placas adyacentes son, sin embargo, muy móviles y se desplazan con facilidad si aplicamos un esfuerzo. A consecuencia de ello, una muestra puede deformarse, no por deslizamiento de placas adyacentes, que es el mecanismo corriente de la deformación plástica, sino por crecimiento y contracción compensada de placas adyacentes.

En una aleación que posea una fase beta capaz de dar martensita cuando se la somete a un esfuerzo puede observarse una propiedad elástica poco común denominada pseudoelasticidad o superelasticidad. En una aleación típica que detente esta propiedad, el metal presenta un comportamiento elástico normal, al ser sometido a un esfuerzo (es decir, se alarga en determinada dirección), hasta que se alcanza un esfuerzo crítico para el cual empiezan a formarse las placas de martensita. Si el esfuerzo aumenta, la muestra continúa alargándose, como si se deformara plásticamente, pero cuando se elimina el esfuerzo, las placas de martensita formadas pasan de nuevo a la fase beta, y la muestra se contrae hasta adquirir su dimensión primitiva y sin mostrar ningún tipo de deformación permanente.

Cuando bajamos la temperatura de una aleación tipo marmem hasta el valor crítico en el cual empieza a formarse martensita, la transformación prosigue su curso espontáneamente. Cuando aumentamos la temperatura por encima del valor crítico, puede formarse aún martensita, pero sólo si se incrementa el valor del esfuerzo aplicado. En el intervalo de temperatura en que se produce la formación de martensita bajo la influencia de un esfuerzo, la elasticidad del material puede superar la elasticidad de las aleaciones ordinarias, según un factor igual o superior a 10, siendo entonces el material superelástico. Aunque la superelasticidad no constituye la característica principal del efecto de memoria de la forma en las aleaciones tipo mar-

a

b

PLANO (110)

FASE DE ELEVADA TEMPERATURA
(FASE BETA)

SUPER-RED B_2

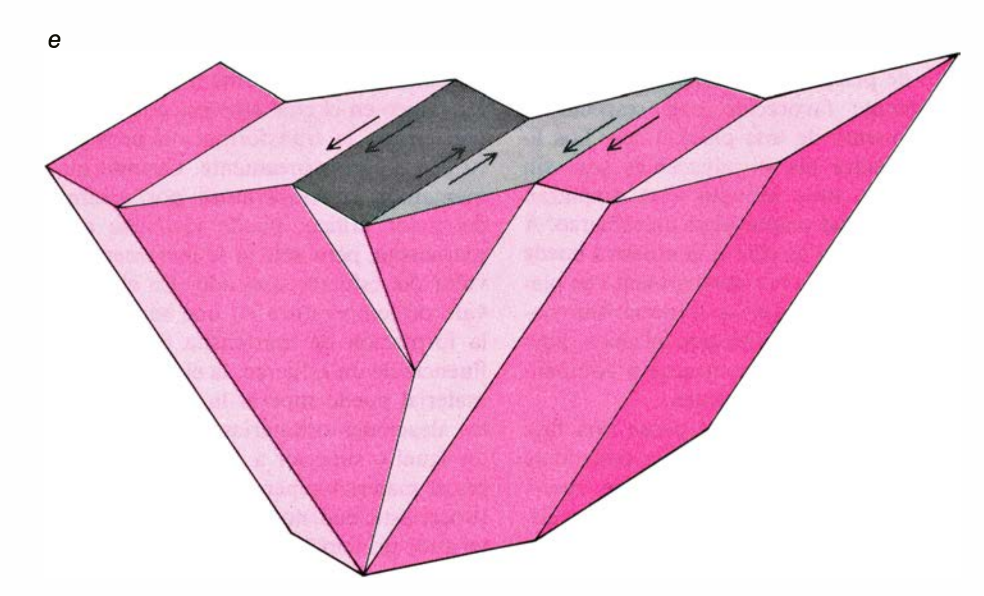
d MARTENSITA ROMBICA

c APILAMIENTO DE PLANOS (110) FORMANDO UNA SUPER-RED B_2

mem, proporciona una cantidad adicional de memoria de la configuración (o fuerza equivalente) que puede aprovecharse para fines específicos. Así pues, una aleación que ya haya sido transformada en martensita presentará deformación superelástica como resultado del movimiento reversible de los límites entre placas de martensita. La reorientación de las placas hará que una variante de martensita crezca a expensas de otra; y ello dependerá de la dirección del esfuerzo con respecto a la energía de deformación interna asociada con cada uno de los mencionados límites.

La capacidad de recibir instrucciones, una cualidad de las aleaciones tipo marmem indicada al principio, proporciona a estos materiales una memoria para dos configuraciones diferentes, es decir, un efecto de memoria "bidireccional". La preparación adecuada del material se consigue limitando el número de variantes de martensita formadas cuando una aleación se calienta y se enfría, alternativamente, por debajo de una temperatura crítica. Esta limitación inhibe la autoadaptación de las placas de martensita e incrementa el valor de la deformación interna. El número de variantes puede limitarse sometiendo la muestra a un esfuerzo mientras se está enfriando desde la temperatura más elevada de la fase beta hasta la temperatura crítica. El esfuerzo favorece la formación inicial de ciertas variantes específicas de martensita, del mismo modo que el esfuerzo a temperatura constante favorece el crecimiento de una variante a expensas de otra. De ahí que pueda prepararse una estructura repitiendo muchas veces la secuencia siguiente, a saber: "betatización" (calentando la aleación hasta que se convierta en la fase beta), enfriamiento rápido, deformación y, de nuevo, betati-

e



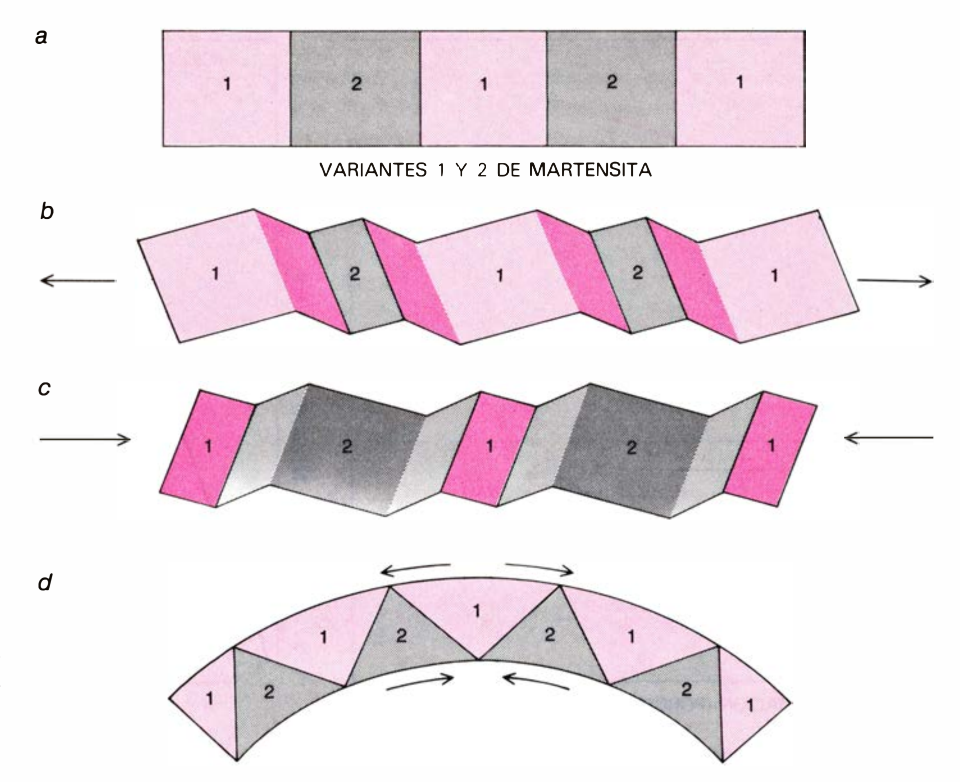
CRISTALES ROMBICOS de martensita. Se forman de un modo espontáneo en una gran variedad de aleaciones cuando se enfrían rápidamente a partir de una temperatura elevada. En la fase beta, o de elevada temperatura, las aleaciones con memoria de la forma adquieren, generalmente, la configuración de cristales cúbicos centrados en el interior y desordenados (*a*). Cuando bajamos la temperatura, una de las especies atómicas (*en color*) de la aleación se sitúa en un lugar preferente dentro de la celda cristalina. Una estructura de este tipo es una superred B_2 (*b*). La transición a martensita puede comprenderse mejor considerando un conjunto de cristales de la superred B_2, donde los planos (110) están apilados en horizontal. Un apilamiento de este tipo, constituido por 10 planos (110), compuestos alternativamente por ocho y siete átomos, puede observarse en *c*. Al enfriarse, la disposición cúbica de la superred B_2 es sustituida por la estructura rómbica de la martensita (*d*). La numeración de los átomos ayuda a clarificar la naturaleza de la recombinación. En una aleación ternaria constituida por cobre, cinc y aluminio, la fase beta produce distintas variantes de martensita que están desplazadas entre sí ángulos de 60 grados (*e*).

zación. Al final se produce espontáneamente un cambio de forma; este cambio sucede en una dirección, a una temperatura que corresponde a la fase beta y, en otra dirección, a una temperatura inferior a la específica de la transformación en martensita.

Voy a resumir, ahora, las condiciones inherentes a la consecución del efecto de memoria de la forma. Me ocuparé, por tanto, de las variables que deben controlarse en el caso de que se desee fabricar algún mecanismo útil. Para que una aleación posea memoria ha de sufrir una transformación a la fase martensita. La temperatura de adquisición de la memoria es función de la temperatura a la que comienza la formación de martensita, o de la temperatura más elevada a la que la aleación pasa a la fase primaria (generalmente la fase beta). La temperatura de la transformación a martensita depende de la composición de la aleación, pero puede modificarse mediante la aplicación de un esfuerzo. Si se frena la recuperación de la forma, se dispondrá de una fuerza proporcional, apta para realizar un trabajo o apretar otro objeto. Para conseguir una recuperación de la forma del 100 por cien, la deformación de una parte debe limitarse a una deformación interna comprendida entre el 3 y el 9 por ciento, según el tipo de aleación de que se trate. Por último, la memoria puede funcionar en un sentido o, mediante preparación adecuada, en dos sentidos.

Los distintos usos de las aleaciones tipo marmem han ido aumentando constantemente. A pesar de su elevado coste, las primeras aleaciones de nitinol se utilizaron para múltiples cometidos en las naves espaciales. Una de las primeras aplicaciones consistió en un mecanismo de bloqueo destinado a un satélite británico en el que un tubo de torsión de nitinol provocaba la liberación de tres paneles de instrumentos. El desprendimiento tenía que ser rápido y seguro para evitar cualquier desequilibrio dinámico que pudiera afectar la estabilidad del propio satélite en rotación.

En otra aplicación inicial, las aleaciones de nitinol solucionaron el problema de la conexión de los conductos hidráulicos en el avión de combate a reacción F-14 fabricado por la Grumman Aerospace Corporation. Los ingenieros de la Grumman dudaban buscando una alternativa a la difícil tarea de soldar conductos que pasan cerca del de la cubierta exterior de aluminio del avión. La Raychem Corporation, con gran experiencia en el campo de los plásticos capaces de contraerse con el calor, propuso una conexión en la que una aleación de nitinol

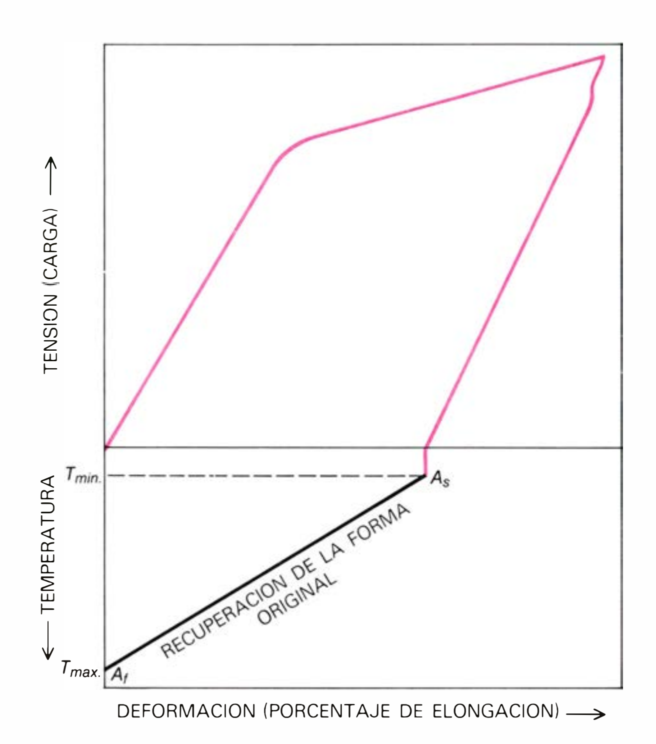


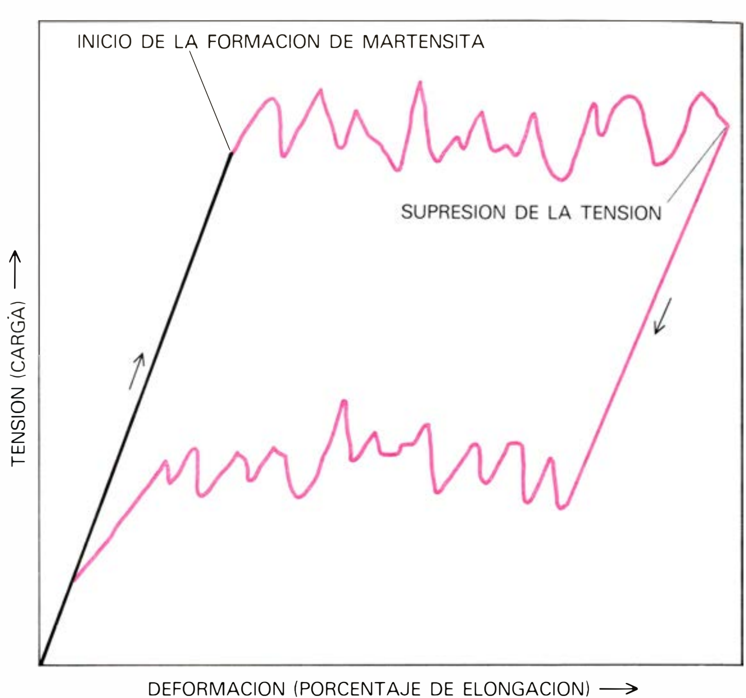
VARIANTES DE MARTENSITA que crecen en proporciones distintas cuando se las somete a un esfuerzo. Para simplificar, aquí sólo consideraremos dos variantes. En ausencia de esfuerzo (*a*) las placas de las dos variantes se desarrollan con igual probabilidad. Si se somete la muestra a un esfuerzo de tensión (*b*), las placas de la variante 1 crecen a expensas de las de la variante 2. Si se somete la muestra a una compresión (*c*), se invierte la preferencia: la variante 2 crece a expensas de la 1. Si la muestra a una compresión (*c*), se invierte la preferencia: la variante 2 crece a expensas de la 1. Si se dobla (*d*), las variantes crecen según los cristales estén sometidos a tensión o a compresión.

montada en caliente se hacía encajar perfectamente al elevarse su temperatura por encima de la de transformación en martensita. La conexión debía mantenerse completamente ajustada hasta temperaturas de menos 120 grados Celsius, de modo que la aleación escogida debía tener una temperatura de transformación situada en la región criogénica.

La temperatura de transformación de la familia de aleaciones de nitinol puede manipularse en un intervalo notablemente amplio, desde menos 273 hasta más 100 grados Celsius, alterando las proporciones relativas de níquel y titanio y añadiendo pequeñas cantidades de otros elementos. Para la conexión hidráulica se encontró una aleación cuya temperatura de transformación era inferior a los menos 120 grados Celsius. La conexión, en forma de manguito, se fabrica a temperatura ambiente de suerte que tenga un diámetro interior aproximadamente un 4 por ciento menor que el diámetro exterior de los tubos que debe unir. Se enfría luego la conexión por debajo de la temperatura de formación de la martensita. Mientras se encuentra a esta baja temperatura se le fuerza a expansionarse, hasta que tenga un diámetro un 4 por ciento mayor que el diámetro del tubo, aplicando una deformación interna global de alrededor de un 8 por ciento. Todavía a una temperatura criogénica, y al objeto de que se mantenga la fase martensita, se encaja el manguito en las terminaciones de los tubos que han de acoplarse. Cuando se calienta a la temperatura ambiente, la fase martensita es reemplazada por la primaria, provocando la contracción del manguito mediante fuerte presión sobre las terminaciones de los tubos. Podemos conseguir un cierre hermético todavía más eficaz si insertamos unas nerviaciones circulares en la superficie interior del manguito. En la actualidad se están ensayando conexiones similares, pero de mayores dimensiones, con el fin de unir conducciones de aire e hidráulicas en los buques mercantes.

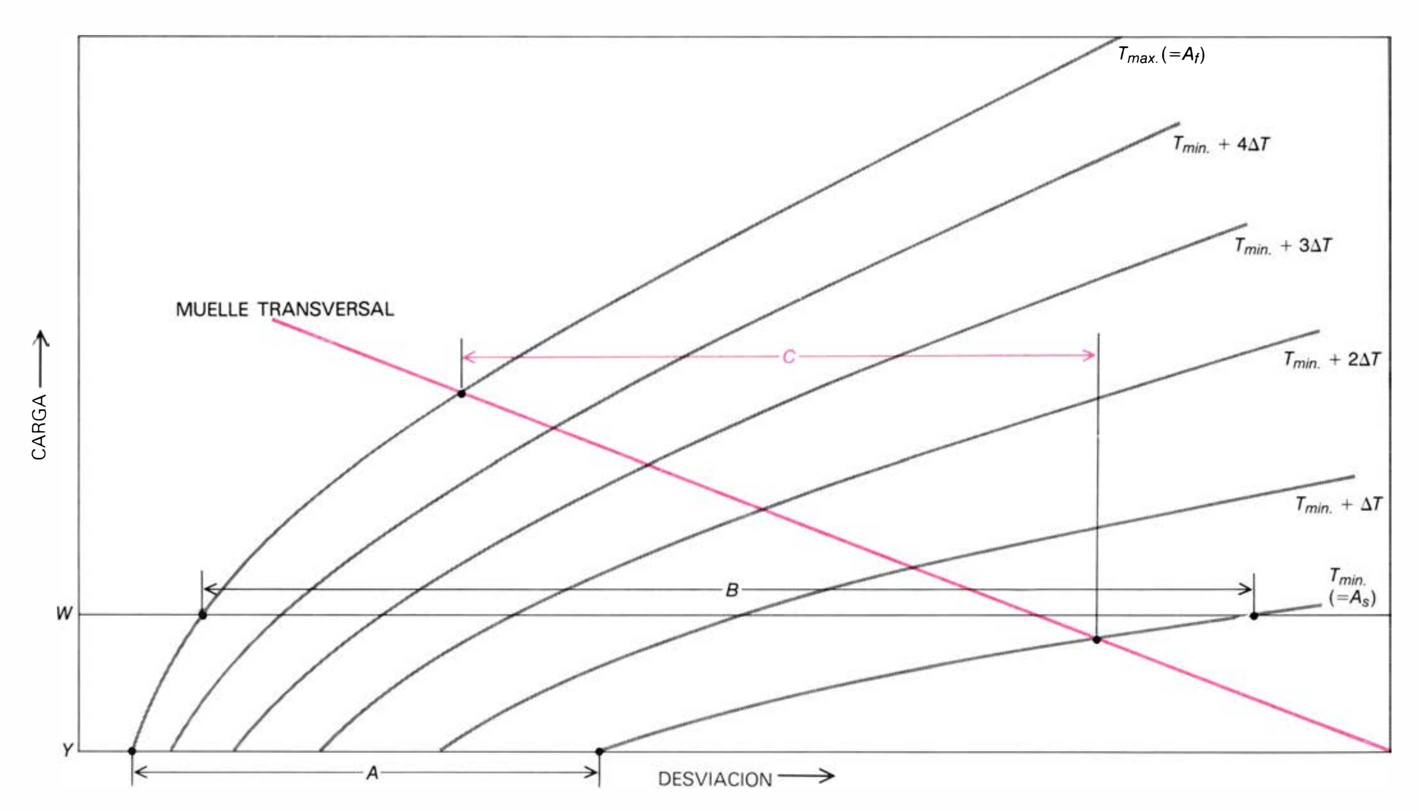
Los ingenieros de la Raychem Corporation también han ideado una conexión eléctrica, que queda ajustada a alta presión, si bien puede desconectarse y volverse a conectar con rapidez. Fabricado a partir de una aleación de nitinol, este mecanismo tiene forma de anillo y se ciñe alrededor de una clavija de conexión. La conexión se abre enfriándola con un chorro de fluorcarbono gaseoso frío, contenido en un recipiente para aerosoles; se cierra cuando se calienta. Las fuerzas ejercidas por tales mecanismos de conexión que poseen la propiedad de





EFECTO DE MEMORIA DE LA FORMA representado en una curva de esfuerzo-deformación que incluye también la influencia de la temperatura. La curva situada por encima de la línea horizontal muestra la respuesta a la deformación de una aleación con memoria de la forma a una temperatura que conserva la fase martensita. Si se aplica un esfuerzo, la muestra se alarga. Si se elimina el esfuerzo, se mantiene una deformación considerable. Si calentamos ahora la muestra hasta una temperatura A_s (o T_{min}), la austenita empieza a sustituir a la martensita. Al aumentar la temperatura, la deformación disminuye rápidamente. En A_f ($T_{máx}$) la conversión en austenita es completa y la muestra adquiere de nuevo su forma original.

SUPERELASTICIDAD, o pseudoelasticidad: es una propiedad de determinadas aleaciones en las que puede inducirse la transformación en martensita en virtud de la aplicación de un esfuerzo. Estas aleaciones presentan una curva típica de esfuerzo-deformación hasta el punto en que, si se sigue aplicando el esfuerzo, empiezan a formarse placas de martensita. Luego, la muestra se alarga plásticamente como si sufriera una deformación permanente. Con la supresión del esfuerzo, las placas de martensita vuelven a pasar a la fase beta y la muestra torna a adquirir su forma original. El alargamiento elástico que resulta de la transformación en martensita puede superar la elasticidad de las aleaciones ordinarias en un factor 10 o más.



CURVAS CARGA-DESVIACION que muestran la respuesta del elemento con memoria de la forma a distintas temperaturas, con y sin un muelle transversal que se oponga a la desviación. Sin carga (*Y*) y sin muelle transversal, se produce una desviación, *A*, con un cambio de temperatura desde $T_{mín}$ hasta $T_{máx}$. En $T_{mín}$ la austenita empieza a sustituir a la martensita y, en $T_{máx}$ la sustitución ya es absoluta. Bajo una carga pequeña, W, el mismo cambio de temperatura produce una desviación mucho mayor, *B*. Si a la desviación se opone un muelle transversal (*en color*), podemos variar la desviación, como respuesta a la carga y a la temperatura, hasta un valor total *C*. El objeto de este muelle transversal es "condicionar" al elemento para que trabaje en aquella parte del espectro carga-desviación-temperatura en el que se puedan aprovechar cantidades importantes de energía.

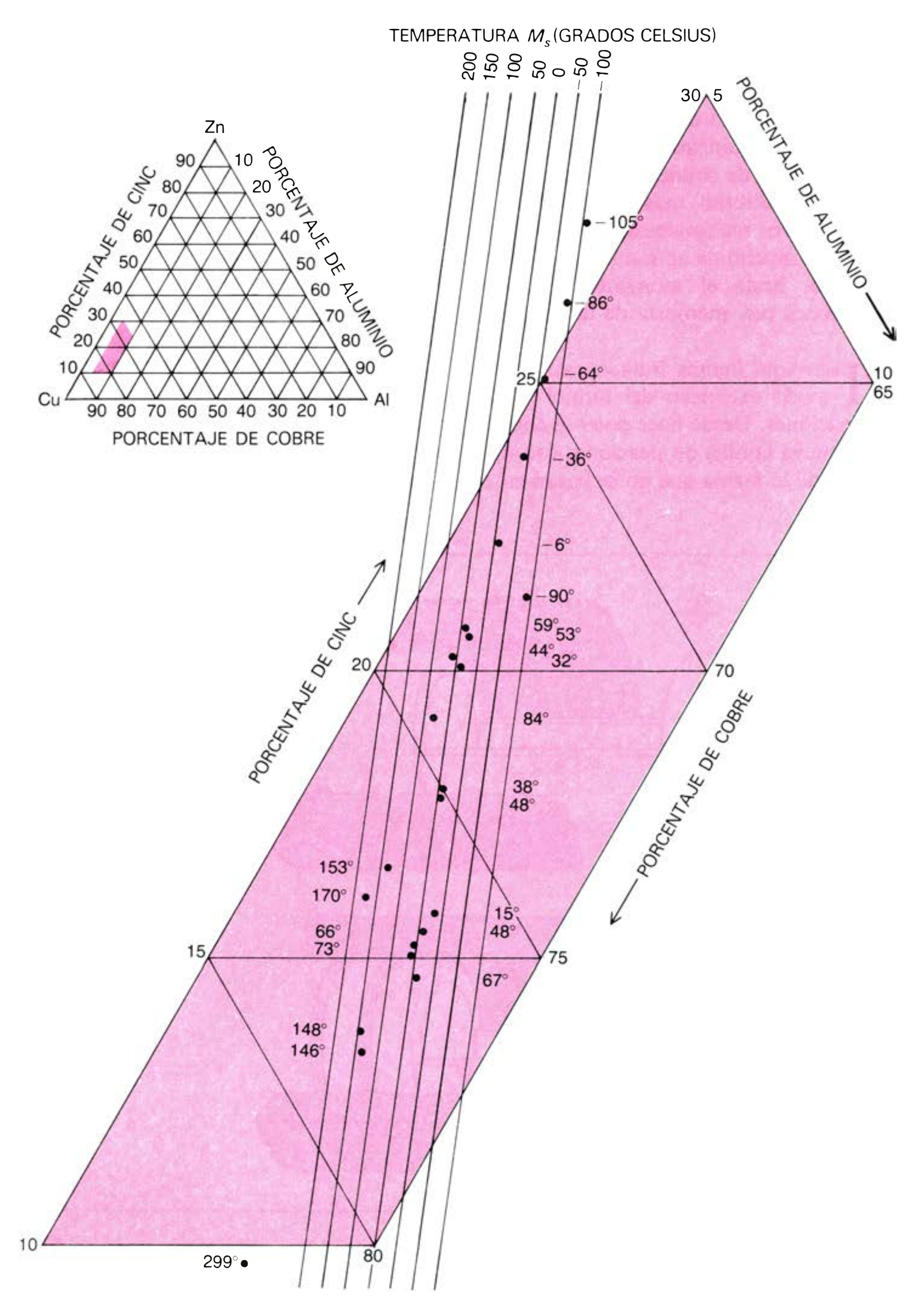
memoria de la forma, son unas 200 veces superiores a las que se desarrollarán mediante la expansión y contracción de un elemento bimetálico del mismo peso. Además, la memoria de la forma tiene lugar a una temperatura determinada y no a lo largo de un amplio intervalo de temperatura, tal como ocurría en el caso de que el mecanismo en cuestión dependiera de la expansión térmica.

En medicina se están investigando diversas aplicaciones prometedoras de las aleaciones de nitinol. Las articulaciones artificiales de miembros son cada vez más frecuentes, pero presentan numerosos problemas. En una articulación grande como la de la cadera, la cavidad ilíaca y la cabeza del fémur artificiales se unen por lo general al hueso mediante un cemento con los consiguientes problemas de falta de alineamiento e incluso de fractura ósea. Otro método de unión utiliza "mariposas" de nitinol unida a la parte de articulación artificial introducida en la cavidad central del hueso. Las mariposas se insertan en frío y se expanden al alcanzar la temperatura del cuerpo, quedando apretadas fuertemente. James Hugues y sus colaboradores, del Mississippi Methodist Rehabilitation Center, de Jackson, han demostrado que el nitinol no es rechazado por los tejidos. Han emprendido, además, una serie de estudios a largo plazo experimentando con animales.

Otro problema ortopédico es el de estirar las partes de un hueso fracturado para asegurar su alineamiento y facilitar su rápida calcificación. El método más corriente de fijación por compresión utiliza clavijas, tornillos y placas; también es lento. Con placas de nitinol, que se insertan enfriadas, el propio calor del cuerpo estira las partes fracturadas del hueso alineándolas. Al estudio de dicha técnica se están dedicando Alan A. Johnson, de la Universidad de Louisville, y Frank P. Alicandri, del Instituto Politécnico de Nueva York.

Una aplicación médica, pero de otro orden, del efecto de memoria de la forma tiene como objeto filtrar los coágulos de sangre del sistema circulatorio antes de que puedan causar graves trastornos. La mayoría de los coágulos de dimensiones peligrosas se forman en las piernas y en el bajo tronco y circulan por las venas hasta el corazón y pulmones donde pueden bloquear un vaso sanguíneo vital. Si bien las substancias anticoagulantes son extraordinariamente útiles, no por eso dejan de presentar sus propios riesgos. Morris Simon, del Beth Israel Hospital, de Boston, y de la Facultad de Medicina de Harvard, concibió la idea de fabricar un filtro en forma de tamiz, con un tamaño de malla de unos dos milímetros, hecho a partir de un sólo trozo de alambre de nitinol. El alambre puede disponerse en forma alargada cuando se enfría por debajo de la temperatura de transformación en martensita. Se ha escogido esta temperatura al objeto de que se halle muy por debajo de la temperatura del cuerpo humano. Cuando se enfría el alambre, para mantener su disposición alargada puede introducirse, mediante un catéter insertado en una vena del brazo, hasta la vena cava, que es el gran vaso sanguíneo que desemboca en el corazón. A medida que el alambre se va calentando, adquiere la forma de tamiz. Las experiencias realizadas con perros han resultado alentadoras.

A los ingenieros les interesa eliminar la parte débil de los mecanismos que deben actuar de forma segura, sin que se tenga que estar siempre, o a intervalos, pendientes de ellos. Un ejemplo que hace al caso es el mecanismo de inscripción existente en decenas de millones de instrumentos de registro y de control industriales. En los instrumentos norma-



TEMPERATURAS DE FORMACION de la martensita, M_s, para el caso de aleaciones ternarias de cobre, cinc y aluminio. Estas temperaturas se pueden hacer variar en mas de 400 grados Celsius con sólo introducir pequeños cambios de composición. Todas las aleaciones con memoria de la forma se sitúan en el extremo con mayor riqueza de cobre (*en color*) del triángulo que representa las mezclas ternarias. La proporción de aluminio varía desde alrededor de un cuatro hasta un 10 por ciento, la de cinc desde algo menos de un 10 hasta un 28 por ciento. En cada caso el resto correspondiente es cobre.

les el brazo de inscripción se mueve por la acción de un galvanómetro, venerable aparato electromecánico que responde a la presencia de una corriente eléctrica.

En su búsqueda de un mecanismo más simple y robusto, los ingenieros de la Foxboro Company han diseñado un mecanismo de inscripción que aprovecha la respuesta de un alambre de nitinol a la memoria de la forma. Bajo control de la tensión, el alambre se alarga y se acorta respondiendo a la cantidad de calor suministrado por una pequeña bobina de inducción. A su vez, la bobina actúa en función de la entrada de potencial en el registrador suministrada por un transductor conectado a lo que el registrador está midiendo. El alambre de nitinol ejerce fuerzas mucho mayores que el dispositivo del galvanómetro, por lo que el mecanismo requiere un número menor de cojinetes y de ejes. Hoy en día se utilizan más de 600.000 de estos nuevos mecanismos, que representan, con mucho, la aplicación más generalizada, hasta el momento, de estas aleaciones con memoria de la forma.

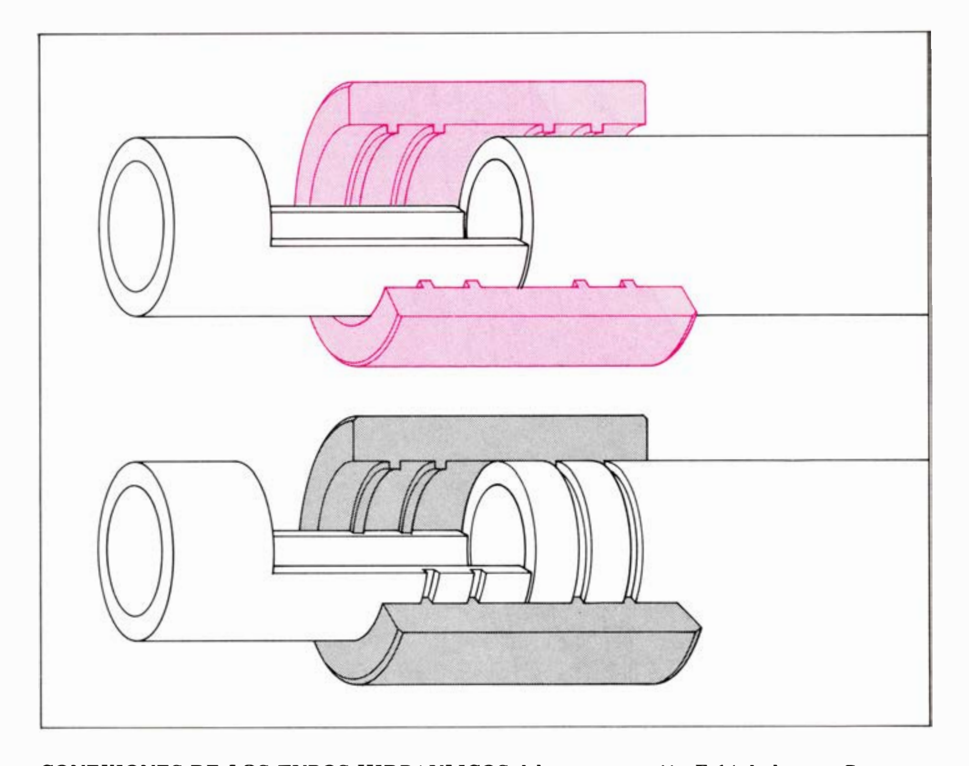
Hasta aquí hemos tratado casi de un modo exclusivo del nitinol y sus aplicaciones. Desde hace poco se conoce una nueva familia de aleaciones con memoria de la forma que no se basan en el níquel y el titanio, sino que utilizan únicamente cobre, cinc y aluminio. Puesto que estas aleaciones ternarias (de tres elementos) son mucho más baratas que el nitinol, y mucho más fáciles de trabajar y fabricar, parecen destinadas a una extensa explotación. Fueron desarrolladas gracias a los trabajos realizados en la Universidad de Lovaina, en la Delta Metal Co. Ltd. y el Fulmer Research Institute Ltd., Gran Bretaña, en colaboración con la Raychem Corporation.

Las nuevas aleaciones ternarias están compuestas por entre un 68 y un 80 por ciento de cobre. El restante 32 a 20 por ciento está repartido entre cinc y aluminio en proporciones variables. Lo notable de estas aleaciones es que bastan pequeñas diferencias de composición para originar grandes cambios en la transformación en martensita, con valores de menos 105 grados Celsius, para los más bajos, hasta valores de 299 grados, para los más elevados. Sin embargo, las aleaciones que pueden fabricarse con facilidad tienen un intervalo menor: está comprendido aproximadamente entre menos 100 y más 100 grados. Para una aleación dada, el efecto de memoria de la forma comprende un intervalo de unos 80 grados. A temperaturas más elevadas, la martensita tiende a ser inestable, de modo que la temperatura superior más corriente a la que se puede operar corresponde a 105 grados. Algunas de las aplicaciones del nitinol interesan también en el campo de las aleaciones ternarias basadas en el cobre: conexiones, muelles y mecanismos de transmisión.

En la búsqueda de nuevas aplicaciones, la compañía Delta ha desarrollado una serie de mecanismos en los que la aleación con memoria de la forma sirve como sensor de temperatura y como transmisor de movimiento. Un mecanismo de este tipo es una válvula termostato utilizable en el radiador del sistema de calefacción de una vivienda o de una oficina. La temperatura a la que se abre la válvula puede determinarse girando un botón que hace variar la compresión de un muelle que, a su vez, actúa sobre otro muelle hecho de la aleación con memoria. La particularidad del muelle con memoria consiste en que la fuerza que ejerce aumenta a medida que se eleva la temperatura. En términos cristalográficos puede decirse que, a medida que la temperatura aumenta, una parte cada vez mayor de martensita se transforma en la fase primaria, la cual está asociada con la memoria del muelle en su posición de expansión.

Un muelle similar con memoria puede regular el sistema de embrague que conecta o desconecta el ventilador del radiador en el sistema de refrigeración de un automóvil. Los mecanismos actuales operan mediante un embrague fluido-viscoso controlado por un muelle bimetálico. El embrague de aleación con memoria elimina los cierres herméticos de los fluidos y proporciona un control de velocidad más flexible. Los ingenieros de la industria automovilística están investigando la posibilidad de recurrir a los sistemas con memoria de la forma para sustituir pequeños motores eléctricos, válvulas accionadas por solenoides y otros diversos mecanismos de control de motores. Una aplicación prometedora es un carburador de paso variable o de inyección variable que proporcione una economía de carburante óptima a lo largo de un intervalo de temperaturas del aire y del carburante.

Toda aplicación que requiera mecanismos con memoria de la forma debe tener en cuenta: la temperatura de funcionamiento, la fuerza y la desviación deseadas y si el movimiento ha de ser lineal o de torsión. Trabajando con una serie de tablas y gráficos confeccionados por la compañía Delta, el ingeniero de proyectos puede determinar, para cada aleación particular, el valor de la fuerza de torsión, o compensadora, necesaria para que el mecanismo con memoria de la forma opere a una temperatura espe-



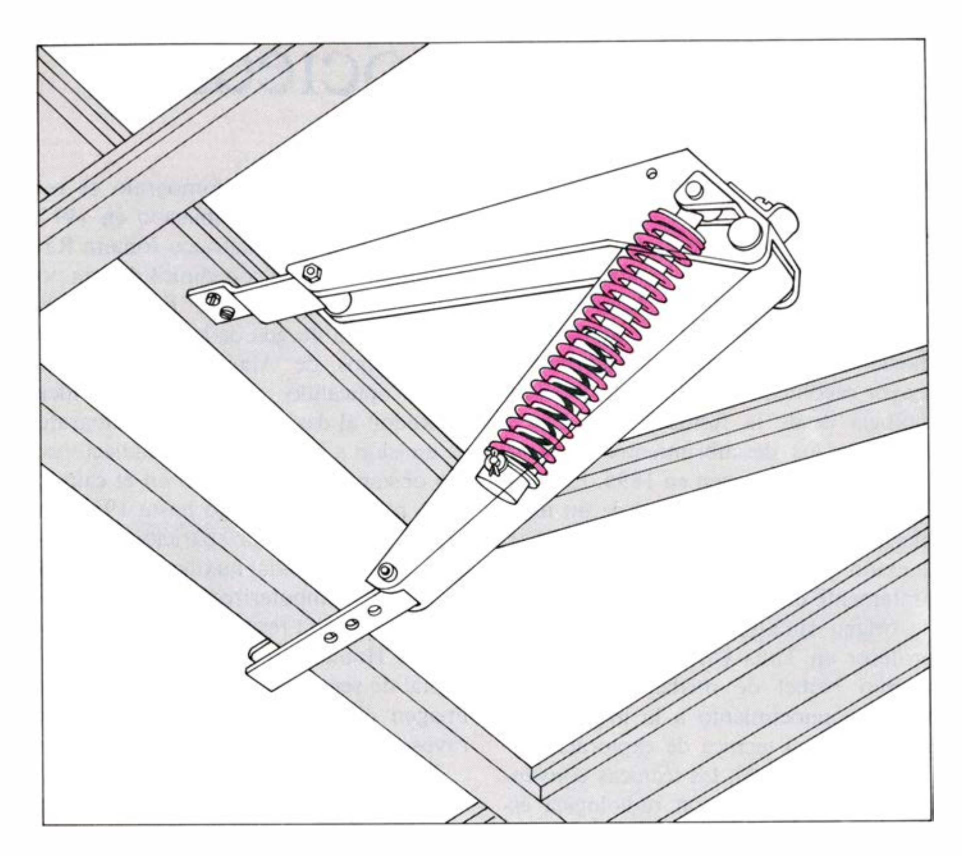
CONEXIONES DE LOS TUBOS HIDRAULICOS del caza a reacción F-14 de la casa Grumman, fabricadas con una aleación de nitinol, cuya temperatura de formación de martensita (M_s) se sitúa en la región criogénica, por debajo los −120 grados Celsius. El manguito de conexión se termina a máquina, a temperatura ambiente, para que tenga un diámetro interior un 4 por ciento menor que el de los tubos que debe unir. Se enfría luego dicho manguito por debajo de la temperatura M_s y se le expansiona mecánicamente para que tenga un diámetro interior un 4 por ciento superior al diámetro exterior de los tubos (*arriba*). Manteniéndolo a una temperatura inferior a M_s se le coloca alrededor de las terminaciones de los tubos. Cuando se calienta la conexión, se produce una contraccion que forma un cierre hermético (*abajo*). Las nerviaciones adicionales hacen más eficaz el cierre hermético al "morder" los tubos.

cífica y sea capaz de proporcionar la fuerza asignada. El número de mecanismos propuestos en patentes se acerca en la actualidad a los cien y la cifra crece día a día.

Vale la pena considerar una aplicación final del efecto de memoria de la forma. Me refiero a la recuperación de energía a partir de calor no aprovechado, existente en el agua a temperaturas bajas, mediante la conversión directa del calor en energía mecánica. La posibilidad de un motor de estado sólido de este tipo empezó a tomarse en consideración poco después del descubrimiento del efecto de memoria de la forma en las aleaciones de nitinol. A la primera patente presentada por Buelher y David Goldstein para un motor con memoria de la forma en 1966 ha seguido lo que hoy en día parece ser una procesión incesante de mecanismos energéticos. El primer motor que demostraba realmente la posibilidad de generar cantidades útiles de energía fue el ideado por Ridgway Banks, del Lawrence Radiation Laboratory, California. Quizá sea porque Banks posee la carrera de música, pero lo cierto es que los mecanismos ideados por él nos proporcionan una sensación de placer estético parecida a la que provocan los instrumentos de cuerda.

El rendimiento teórico de un motor de estado sólido puede analizarse si se tienen en cuenta el calor latente de transformación, el esfuerzo suplementario producido por un cambio adicional de temperatura y la diferencia en energía libre entre la fase de martensita y la fase de elevada temperatura. Mediante un análisis de este tipo, L. Delaey, de la Universidad de Lovaina, ha encontrado un rendimiento teórico de un 4 o 5 por ciento, que, para un motor que trabaja con la modesta diferencia de temperatura equivalente a 20 grados Celsius, corresponde a una quinta parte del valor standard del rendimiento térmico del ciclo de Carnot.

El primer motor de Banks operaba mediante unas vueltas de alambre de nitinol que subían y bajaban sobre unos radios unidos mediante un cigüeñal a una rueda. Cuando la rueda giraba, las vueltas anulares del alambre se bañaban primero en agua fría y luego en agua caliente, haciendo que esas vueltas se abrieran y cerraran. El mecanismo se parece al motor radial de varios pistones. En otro sistema dispositivo, una serie de bandas de una aleación ternaria con memoria se abren y cierran como un acordeón al calentarse y enfriarse, produciendo un movimiento axial que un cigüeñal puede convertir en rotacional al igual que en el motor de Banks.



MECANISMO SIMPLE DE APERTURA de una ventana que podría aplicarse a un invernadero. Se acciona por un muelle hecho de una aleación rica en cobre y con memoria de la forma. A temperaturas por debajo de unos 18 grados Celsius, el muelle se halla contraído y la ventana permanece cerrada. Cuando la temperatura se eleva por encima de 18 grados, el muelle con memoria de la forma vence la fuerza de un muelle transversal y empieza a abrir la ventana. A 25 grados, el muelle se halla extendido.

Uno de los últimos motores presentados por Banks utiliza el principio del balancín móvil descubierto por el ingeniero inglés del siglo XVIII Thomas Newcomen. El "pistón" consta de 88 hilos de alambre de nitinol de unos 40 centímetros de longitud. Cuando el mecanismo constituido por el alambre se sumerge alternativamente en agua fría y caliente, su contracción y expansión se transmiten, mediante el balancín móvil y las uniones transversales, a una rueda de salida. Un cigüeñal secundario (llamado a veces charleston, por su parecido con el baile popular de los años 1920) controla el tiempo de permanencia en los dos baños de agua, de la misma manera que una conexión variable controla la acción de la válvula en un motor de vapor.

Dado el bajo rendimiento del sistema, cabe preguntarse qué tipos de fuentes podrían suministrar baja energía térmica en cantidades suficientes para generar energía útil. Una de estas fuentes es, evidentemente, el océano, en especial en los trópicos, donde la diferencia de temperatura a lo largo del año entre la superficie oceánica calentada por el sol y el agua fría profunda es, por lo menos, de 25 grados Celsius. Otra fuente potencial importante de baja energía térmica ha sido indicada por J. S. Cory. Este observó que detrás de una presa hidroeléctrica que detrás de toda presa hidroeléctrica elevada existe un gradiente de temperatura importante, entre la región superficial del embalse y su fondo, que puede variar desde unos grados en los meses de invierno hasta 18 grados en verano. Lo sorprendente es que la energía térmica de tan importante masa de agua es mucho mayor que la energía potencial propia fuera de gravedad. La capacidad para extraer esta energía, de forma parecida a lo que se ha dicho respecto a la diferencia de temperatura en el océano, exige la utilización de algún tipo de motor térmico. El motor con memoria de la forma es un posible candidato.

Como en muchos fenómenos descubiertos anteriormente, el efecto de memoria de la forma es una solución que necesita encontrar problemas por resolver. He mencionado algunas aplicaciones que ya han podido explotarse comercialmente. Entre estas últimas y las aplicaciones más especulativas antes señaladas existen nuevos aparatos que se encuentran hoy en fase experimental. Entre ellos, tenemos desde bombas para extraer petróleo de pozos poco profundos hasta una amplia gama de aparatos de control automático. Y podemos estar seguros de que habrá más.

Ciencia y sociedad

Nobel de medicina

En 1972, como resultado de trabajos emprendidos en 1968, aparece bajo el nombre de EMI-Scanner un sistema no invasor de exploración mediante rayos X que ha supuesto la mayor revolución en la historia de la radiología desde la fundación de ésta a partir de los descubrimientos de Wilhelm Konrad Röntgen en 1895. El EMI-Scanner es el hijo espiritual de un ingeniero de esta firma comercial británica, llamado Godfrey Hounsfield, el cual, juntamente con el físico estadounidense, de origen sudafricano, Alan Cormack, profesor en Tufts University, recibió el premio Nobel de medicina en 1979, como reconocimiento a la importancia de esta nueva técnica de exploración.

Como es sabido, las técnicas convencionales de exploración radiológica están basadas en la utilización de radiaciones lo suficientemente penetrantes como para atravesar el sujeto explorado, pero no tanto que hagan imposible una imagen en la que se pongan de manifiesto los distintos tejidos mediante la distinta absorción de los mismos, siendo esta imagen, generalmente fotográfica, y necesariamente de dos dimensiones, el resultado de superponer sobre un plano único las informaciones procedentes de los distintos planos del sujeto atravesado por la radiación, lo que es causa de que se pierda una buena parte de éstas.

En el "scanner" se emplea el mismo procedimiento de exploración, utilizando un generador y un detector, no fotográfico, de rayos X situados diametralmente opuestos respecto del sujeto explorado, pero siendo generador y detector susceptibles de girar solidariamente alrededor del sujeto, que permanece estacionario. Por su parte, el detector recibe solamente la radiación correspondiente a las trayectorias que atraviesan el plano de interés y a partir de esta información mediante un tratamiento matemático adecuado, realizado en un ordenador, se reconstruye la imagen correspondiente al plano transversal explorado. En otras palabras, se obtiene la imagen correspondiente a un corte transversal, normal al eje de giro del generador y detector, siendo esta la razón por la cual el "scanner" recibe el nombre de tomógrafo. La resolución de la imagen obtenida es de unos 2 milímetros, lo que permite visualizar órganos, tejidos, tumores, hemorragias, etc. con gran precisión y rapidez.

La idea básica del tomógrafo se remonta a un trabajo publicado en 1917 por el matemático austríaco Johann Radon, pero la utilización clínica de ésta no empieza a vislumbrarse hasta el comienzo de la década de los 60, gracias a los trabajos de Alan Cormack y de Kuhl, aplicando este último la idea de Radon al desarrollo de un tomógrafo de emisión a base de fuentes radiactivas. No obstante, la revolución en el campo de la radiología no llegó hasta 1917, en que EMI anunció la aparición del primer tomógrafo axial auxiliado con ordenador ("Computerized axial tomograph", o CAT) resultado de la idea genial de Hounsfield de aplicar el proceso digital de señal a la reconstrucción de la imagen obtenida en un tomógrafo de rayos X.

Para reconstruir la imagen de dos dimensiones correspondiente a un corte transversal es preciso realizar un gran número de medidas de transmisión de la radiación a lo largo de trayectorias comprendidas en el plano de corte. De acuerdo con la fórmula de Beer, si exploramos un cuerpo con un haz de rayos X de intensidad I_0, la intensidad recibida por el detector es igual a

$$I = I_0 \exp \int_{\text{generador}}^{\text{detector}} -\mu(x,y)\, dl$$

donde $\mu\,(x,y)$ es la absorción del sujeto en cada punto y dl un elemento de la trayectoria recorrida por el haz.

El problema de reconstrucción de la imagen del objeto explorado, a partir de una pluralidad de lecturas, consiste en obtener un estimador $\hat{\mu}$ de la función μ lo más exacto posible, no siendo posible otra cosa ya que, por otra parte, la exploración se realiza discretamente.

Existen varios procedimientos para llevar a cabo el proceso de reconstrucción, siendo el más frecuente el conocido como de retroproyección corregida ("filtered back projection"). Este u otro procedimiento adecuado es llevado a cabo en un ordenador y su resultado, la imagen reconstituida, presentado en una pantalla de rayos catódicos y almacenada en discos o cintas magnéticos.

Los tomógrafos de esta primera generación emplean unos 4 minutos para realizar una exploración completa de un corte. Las máquinas de la segunda generación utilizan varios haces de rayos X y una distribución espacial de detectores, con un tiempo de exploración entre 10 segundos y 2 minutos. Los tomógrafos de la tercera generación son aún más rápidos, empleando entre 5 y 10 segundos para realizar una exploración. Estos tomógrafos emplean un haz de forma de abanico y una distribución de detectores compuesta por varios cientos de éstos. (M.A.)

Rojo y azul

La medición de los desplazamientos Doppler presentes en los espectros de los objetos celestes constituye una de las herramientas más potentes de los astrónomos. Un desplazamiento de las líneas espectrales hacia las longitudes de onda mayores, o hacia el extremo rojo del espectro, nos indica que el objeto observado (polvo o gas) se está alejando de la Tierra. Un desplazamiento hacia el extremo azul del espectro indica que la fuente que emite la radiación se está acercando. En la astronomía de grandes distancias predominan los desplazamientos hacia el rojo. Así, el constante incremento del corrimiento hacia el rojo en función de la distancia mostrado por las galaxias indica que el universo está en expansión. Tan sólo algunas galaxias cercanas presentan desplazamientos hacia el azul. Dentro de nuestra propia galaxia, las estrellas presentan una mezcla aleatoria de pequeños corrimientos hacia el rojo y el azul. El desplazamiento Doppler es particularmente útil para detectar sistemas estelares binarios. Cuando el espectro de una determinada fuente aparece desplazado con periodicidad ligeramente hacia el azul y ligeramente hacia el rojo, es verosímil deducir que dicha fuente está moviéndose sobre una órbita, acercándose y alejándose del sistema solar.

Durante los últimos años, los astrónomos se han sentido fascinados por el descubrimiento de un objeto peculiar en nuestra propia galaxia que presenta un desplazamiento hacia el azul más de 100 veces mayor que los anteriores conocidos, de una magnitud comparable, aunque de signo opuesto, a los desplazamientos hacia el rojo de los quasars más próximos, los cuales son objetos extragalácticos que se cree están situados a distancias enormes. El misterio que representa esta nueva fuente se desvela un tanto por el hecho de que el desplazamiento hacia el azul de su espectro viene igualado simultáneamente por un desplazamiento hacia el rojo de la misma magnitud. La fuente en conjunto no se acerca ni se aleja con la velocidad que indicarían cada uno de los corrimientos, al rojo y al azul, por separado. La explicación más plausible para este doble fenómeno es que la fuente sea una estrella de neutrones, probablemente algún tipo

de pulsar, que emita dos haces de partículas a alta velocidad y en direcciones opuestas.

La fuente emisora de los haces se calificó al principio, a mediados de la década de 1960, como una estrella de magnitud 14 muy peculiar. Más tarde se designó por SS433. Se encuentra a unos 11.000 años-luz de la Tierra, en la constelación del Aguila, cuatro grados al norde del ecuador celeste. Durante esta última década, se ha identificado a SS433 como fuente de radioondas y de rayos X. Dentro de un programa encaminado a estudiar las estrellas emisoras de rayos X, Bruce Margon, de la Universidad de California en Los Angeles, y un grupo de seis colegas iniciaron observaciones espectroscópicas de SS433, en agosto de 1978, con el telescopio reflector Shane, de tres metros, situado en el Observatorio Lane. "Los primeros espectros que conseguimos", ha dicho Margon, "nos indicaron que existía algo sumamente raro". En una serie de espectros realizados durante ocho noches consecutivas, una de las líneas de emisión se desplazó hacia el rojo a intervalos prácticamente iguales desde una longitud de onda de 7400 angstrom hasta los 7620, a la vez que una segunda línea se desplazó hacia el azul desde 6120 a 5970 angstrom. Los cambios en longitud de onda corresponden a un incremento de la velocidad a que se aleja la fuente de 9000 kilómetros por segundo y un aumento de la velocidad con que se nos acerca de 7400 kilómetros por segundo. Las líneas consideradas corresponden a líneas de emisión del hidrógeno y del helio atómicos a una temperatura de 10.000 grados Kelvin.

Los estudios posteriores indicaron que las emisiones procedían de haces que giraban en direcciones opuestas en los cuales las partículas alcanzaban una velocidad de 81.000 kilómetros por segundo, el 27 por ciento de la velocidad de la luz. En cada período, de alrededor de 164 días de duración, los haces describen dos figuras cónicas. El eje común de estos dos conos forma un ángulo de 78 grados con la línea de visión desde el sistema solar. El ángulo del vértice de cada cono es de unos 34 grados. Debido a que los haces se ven lateralmente, las velocidades máximas observadas según la línea de visión son mucho menores que las velocidades verdaderas: unos 32.000 kilómetros por segundo hacia el sistema solar para el haz predominantemente "azul", y unos 54.000 kilómetros por segundo alejándose del sistema solar para el haz "rojo".

Durante una pequeña fracción de cada ciclo de 164 días, la dirección a que apunta el haz azul queda más allá de la perpendicular a la línea de visión y, por tanto, se desplaza brevemente hacia el rojo. De igual modo, durante el mismo intervalo de tiempo, el haz rojo apunta a una dirección más próxima que la referida perpendicular. En esta parte de ciclo, donde sería de esperar que el haz rojo se desplazase hacia el azul, resulta que no es así, sino que sigue mostrando un corrimiento hacia el rojo debido a la dilatación relativista del tiempo, lo que da lugar al efecto conocido con el nombre de desplazamiento Doppler transversal. A las velocidades a que se mueven las partículas, aparece un desplazamiento hacia el rojo constante y equivalente a una velocidad de 11.000 kilómetros por segundo. Esta corrección de 11.000 kilómetros por segundo se aplica, como es natural, constantemente a ambos haces y en consecuencia anula el desplazamiento hacia el azul del haz rojo.

Se ha supuesto que SS433 es un sistema estelar binario. Recientemente, David Crampton, Anne Cowley y John Hutchings, del Dominion Astrophysical Observatory, en Canadá, han confirmado esta sospecha refinando las medidas del efecto Doppler. La fuente de luz visible está en órbita alrededor de una estrella demasiado débil para ser fotografiada. Las emisiones de SS433 en el dominio del visible, de los rayos X y de las ondas de radio proceden evidentemente de un disco de acreción alrededor de una estrella de neutrones, probablemente residuo de una supernova. Un tal disco consiste en la materia que se está transfiriendo de la estrella normal del sistema binario al miembro altamente colapsado: la estrella de neutrones. I.S. Shklovsky, del Instituto para la Investigación Espacial de la Academia de Ciencias de la Unión Soviética, ha sugerido que la masa se transfiere a la estrella de neutrones más deprisa de lo que puede ser absorbida. Por algún mecanismo, la estrella de neutrones, quizás un joven pulsar que gira con rapidez, repele parte de la masa que le llega hacia afuera en dos chorros delgados, dando lugar a los haces observados.

El pegamento de los quarks

Entre la multitud de vocablos caprichosos introducidos en la física de las partículas elementales durante las últimas décadas –como "quark", "color" y "encanto" –quizás el escogido con mayor acierto sea "gluon" (del inglés *glue*, que significa cola de pegar). El gluon se ha definido como la entidad que mantiene unidas las partículas denominadas quarks, formando así los protones, neutrones, piones y todas las restantes entidades que se agrupan bajo el nombre genérico de hadrones, o partículas que interaccionan fuertemente. Se cree que el poder de adherencia del gluon es enorme, hasta el punto de que resulta imposible extraer el quark de un hadrón, con independencia de la fuerza que actúe sobre él. Además, también parece que el propio gluon está confinado de un modo permanente: así como los quarks no pueden ser aislados, es imposible destilar una gota del pegamento que los mantiene unidos. A pesar de todo, hay pruebas substanciales de la existencia de los quarks. Ahora ya todo un haz de experimentos han configurado un avance de corroboración de la existencia de los gluones.

El gluon es similar en algunos aspectos al fotón, que es el cuanto de radiación electromagnética. Ambas partículas carecen de masa y se mueven a la velocidad de la luz, y pueden contemplarse como los agentes de sendas fuerzas fundamentales de la naturaleza. El fotón transmite la fuerza electromagnética. Así, podemos concebir la atracción entre el electrón y el protón en un átomo de hidrógeno como el constante intercambio de fotones, emitidos por una partícula y absorbidos por la otra. Sólo las partículas que tienen carga eléctrica pueden emitir o absorber fotones y, en consecuencia, son las únicas partículas que están sometidas a interacciones electromagnéticas.

La fuerza transmitida por el gluon nos es menos familiar que el electromagnetismo y sus efectos son más complicados. Recibe el nombre de fuerza fuerte o fuerza del color, y viene gobernada por la propiedad de los quarks denominada color. (El color de los quarks no tiene nada que ver, naturalmente, con los colores que percibimos a través de la visión.) Mientras que sólo existe una clase de carga electromagnética, que puede tomar valores positivos o negativos, hay tres clases de carga de color, cada una de las cuales tiene también dos posibles valores. Se dice que los quarks tienen los colores rojo, azul y verde; los correspondientes antiquarks son antirrojo, antiazul y antiverde. De acuerdo con un principio fundamental de la teoría de interacciones entre quarks, los hadrones sólo pueden estar formados por ciertas combinaciones de los colores de los quarks, las llamadas combinaciones blancas (o sin color). El protón y otros muchos hadrones con una estructura similar constan de tres quarks, uno de cada color. El pion y las partículas afines están formadas por un quark y un antiquark, con colores que se anulan mutuamente; por ejemplo, el quark puede ser rojo y el antiquark antirrojo.

Al igual que las partículas cargadas pueden mantenerse unidas debido al intercambio de fotones, los quarks en un hadrón están cementados entre sí por el intercambio de gluones. Sin embargo, ahora aparece una complejidad mayor. Aunque el fotón transmite la fuerza electromagnética, es eléctricamente neutro; cuando un electrón emite un fotón, el electrón cede energía y momento, pero no ve alterada su carga eléctrica. El gluon, por su parte, transporta carga de color. En realidad hay ocho especies de gluones, de acuerdo con las distintas combinaciones de colores y anticolores, combinaciones todas en las que el cambio de color no se anula. Por ejemplo, un tipo de gluon tiene los colores rojo y antiazul, otro es verde y antirrojo. Consecuencia de estas cargas de color es que los gluones pueden interaccionar entre sí con la misma facilidad con que lo hacen con los quarks. Otra consecuencia es que, cuando un quark emite un gluon, se modifica el color del quark. Si un quark rojo emite un gluon rojo-antiverde, entonces, a fin de conservar constante la carga de color total de las dos partículas, el primero debe transformarse en un quark verde.

Debido a que los gluones, al igual que los quarks, no pueden observarse directamente, las pruebas de su existencia se han de obtener a partir de las mediciones realizadas sobre los hadrones ordinarios. Los experimentos efectuados en aceleradores con los que se han conseguido esas pruebas comenzaron con el estudio de electrones y sus antipartículas, positrones, almacenados en haces que giraban en sentidos contrarios y que chocaban frontalmente. Los electrones y positrones no son hadrones, no están formados por quarks, pero cuando entran en colisión se aniquilan mutuamente para dar un estado de energía pura a partir del cual puede materializarse un par quark-antiquark. Se precisa que el quark y el antiquark tengan colores opuestos.

Para energías bajas, las colisiones electrón-positrón originan, con mayor frecuencia, un conjunto de partículas que se alejan del punto de impacto en todas direcciones. Al aumentar la energía, menudea otra clase de resultado: aparecen dos haces de partículas, en su mayoría hadrones, que se alejan del punto de impacto en sentidos diametralmente opuestos. La interpretación de estos hechos es muy sencilla: cuando el quark y el antiquark se mueven en direcciones opuestas se crean nuevos pares quark-antiquark en sus proximidades. A continuación, los quarks y los antiquarks se unen entre sí para formar hadrones, algunos de los cuales pueden desintegrarse posteriormente para dar todavía más hadrones y otras partículas. Todas las partículas descendientes del quark (o del antiquark) original continúan moviéndose en aproximadamente la misma dirección y, en consecuencia, forman un chorro muy paralelo.

Se han observado estos chorros de dos puntas en varios laboratorios. A ellos debemos información sobre la estructura de quarks de los hadrones. Los indicios recientes sobre la existencia de los gluones proceden de sucesos todavía más elaborados en los que se han empleado tres chorros.

Cuando un electrón y un positrón colisionan violentamente, el quark y el antiquark se crean con una energía cinética alta. En tal caso es probable que al menos una de esas partículas ceda parte de su energía emitiendo un gluon. (El proceso electromagnético análogo, en donde un electrón de alta energía emite un fotón, recibe el nombre de bremsstrahlung, o radiación de frenado.) La emisión del gluon cambia el color del quark, si bien sigue siendo nulo el color total del quark, el antiquark y el gluon.

La emisión del gluon altera también la geometría del suceso observado. Al igual que el quark y el antiquark iniciales, el gluon inducirá la creación de pares quark-antiquark en el vacío circundante, de forma que dicho gluon se desintegrará para formar hadrones normales. En consecuencia, los productos de desintegración del gluon forman un tercer chorro, y la distribución global de las partículas tiene tres puntas, situadas todas ellas en un mismo plano.

Entre los efectos de las colisiones producidas en PETRA, un moderno anillo de almacenamiento de partículas situado cerca de Hamburgo, se han encontrado evidencias de sucesos de tres puntas. Los electrones y los positrones almacenados en PETRA pueden hacerse colisionar con energías que se aproximan a los 30.000 millones de electronvolt, que es alrededor de tres veces la energía disponible en el mayor de los anillos de almacenamiento precedentes. Sin embargo, gran parte de los sucesos registrados hasta ahora se producen a energías sólo algo por encima del umbral donde comienza a hacerse probable la emisión de gluones. Debido a ello, el número de sucesos de tres puntas resulta comparativamente pequeño, y en la mayoría de ellos el gluon sólo diverge ligeramente del quark que lo ha emitido. Por tanto, el tercer chorro no puede distinguirse con facilidad. Lo que se observa en realidad se parece mucho más a un suceso de dos chorros, con uno más difuso que el otro lo que indica que consta de dos haces de partículas solapados.

Teoría neutralista de la evolución molecular

A nivel molecular, la mayor parte de los cambios evolutivos y de la variabilidad dentro de una especie no se deben a la selección sino a la deriva genética de genes mutantes selectivamente equivalentes

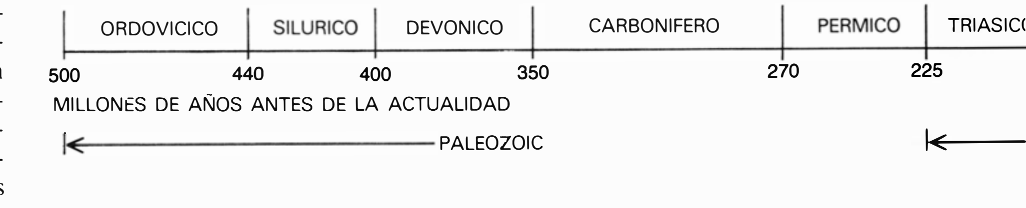
Motoo Kimura

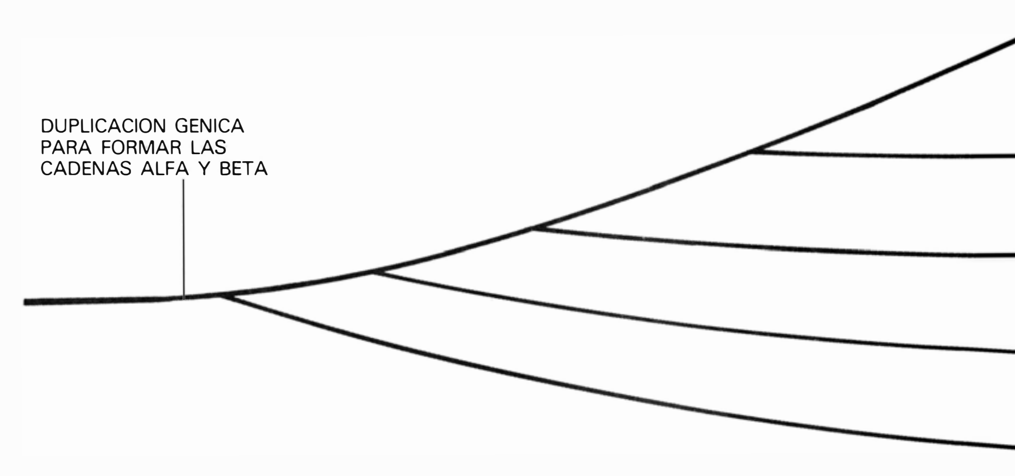
La teoría darwinista de la evolución a través de la selección natural está firmemente establecida entre los biólogos. Dicha teoría sostiene que la evolución resulta de una interacción entre la variación y la selección. En una especie, se produce en cada generación un gran aporte de variación que se debe, por un lado, a la mutación de los genes y, por otro, a la ordenación al azar de los genes en la reproducción. Los individuos cuyos genes den origen a los caracteres mejor adaptados al ambiente serán los más aptos para sobrevivir, reproducirse y dejar supervivientes que se reproducirán a su vez. Las especies evolucionan por acumulación de genes mutantes adaptativos y de los caracteres originados por dichos genes.

Bajo este punto de vista, cualquier alelo mutante (forma mutante de un gen) es más adaptativa (o menos) que el alelo del que deriva. Dicho alelo va aumentando su presencia en la población sólo si logra pasar la rigurosa prueba de la selección natural. Desde hace más de una década vengo defendiendo un punto de vista diferente. En mi opinión, la mayoría de los genes mutantes que sólo se detectan por medio de las técnicas químicas de la genética molecular son selectivamente neutros, es decir, no tienen adaptativamente ni más ni menos ventajas que los genes a los que sustituyen; a nivel molecular, la mayoría de los cambios evolutivos se deben a la "deriva genética" de genes mutantes selectivamente equivalentes.

Evolución del darwinismo

La controversia entre el punto de vista neutralista y el "panseleccionista" tiene su origen en el propio desarrollo de la moderna teoría "sintética" de la evolución. Cuando Darwin formuló su teoría original, no se conocían ni los mecanismos de la herencia ni la naturaleza de las variaciones hereditarias. En nuestro siglo, gracias al desarrollo de la genética mendeliana, quedó abierto el camino para el establecimiento de una base genética de las intuiciones de Darwin. Esto se consiguió, en gran parte, merced al descubrimiento por H. J. Muller de la naturaleza fundamental del gen y por medio de las técnicas de la genética de poblaciones desarrolladas principalmente por R. A. Fisher, J. B. S. Haldane y Sewall Wright. Sobre estos cimientos, los posteriores estudios de poblaciones naturales realizados por Theodosius Dobzhansky, los análisis paleontológicos





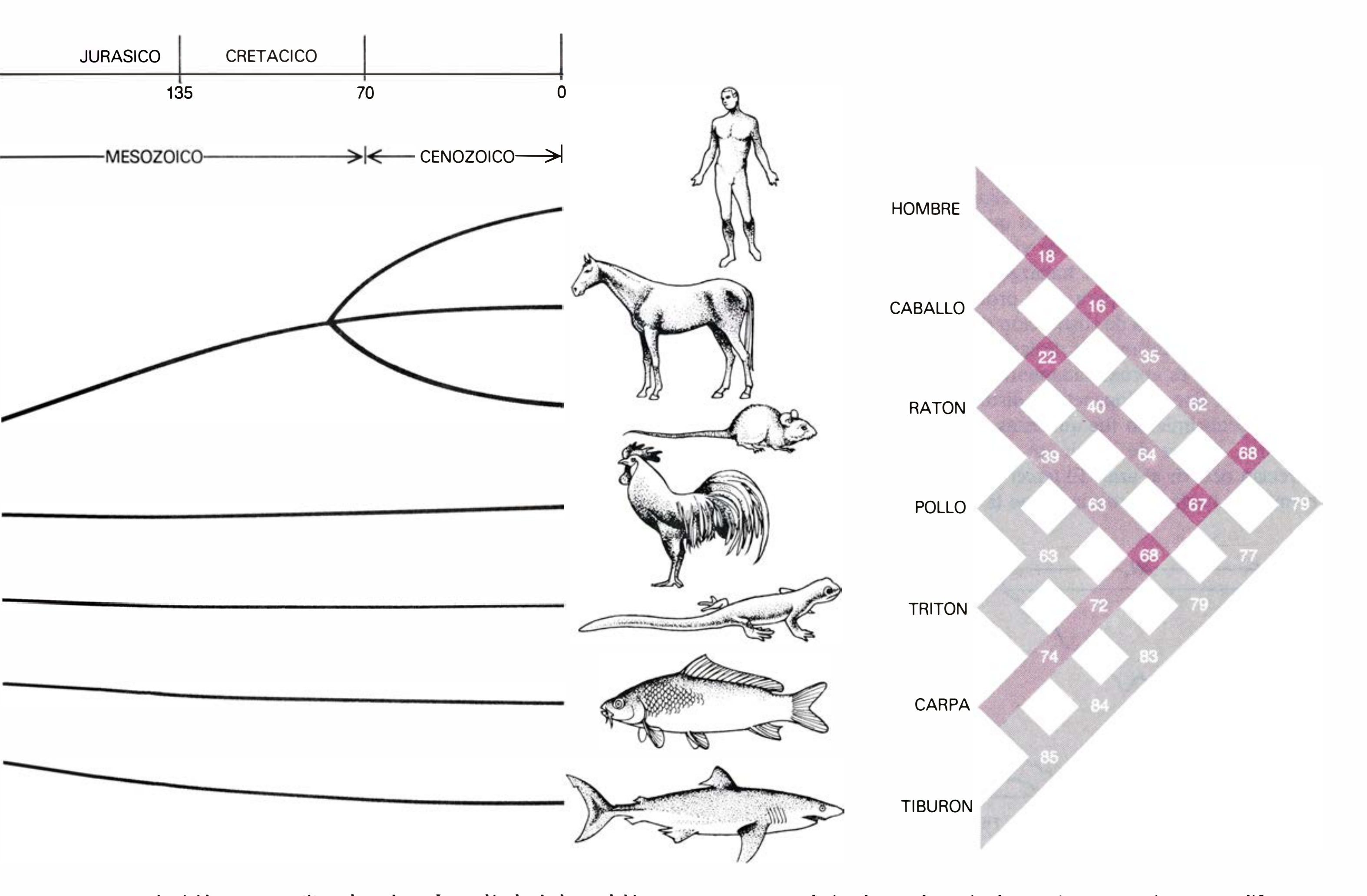
ARBOL FILOGENETICO que pone de manifiesto las relaciones evolutivas entre siete vertebrados distintos y muestra cómo y cuándo divergieron sus líneas filogenéticas en el tiempo geológico. La tabla de la derecha muestra hasta qué punto difiere una misma proteína, la importante cadena alfa de la hemoglobina, en los siete animales; concretamente da el número de diferencias en la secuencia de

de George Gaylord Simpson, la "genética ecológica" de E. B. Ford y su escuela y otras investigaciones fueron construyendo el impresionante edificio de la teoría neodarwinista.

A principio de los años sesenta se aceptaba en general que todo carácter biológico podía interpretarse a la luz de la evolución adaptativa por medio de la selección natural, y que casi no había genes mutantes que fueran selectivamente neutros. Ernst Mayr lo expresó así en 1963, "Considero... sumamente improbable que algún gen pueda permanecer selectivamente neutro por algún período de tiempo". Mucho se ha dicho acerca de cómo interactúan los genes, cómo se organizan los acervos genéticos de las especies y cómo cambian las frecuencias génicas en las poblaciones en el curso de la evolución. Sin embargo, estas conclusiones eran necesariamente deducciones basadas en observaciones a nivel fenotípico: a nivel de la forma y la función resultantes de la actuación de los genes. No había manera de saber lo que realmente ocurría en la evolución a nivel de la estructura interna del gen.

Mientras tanto, la teoría matemática de la genética de poblaciones iba adquiriendo una complejidad progresiva (fenómeno poco frecuente en biología). Particularmente notable fue la estructura teórica a que dio lugar la utilización de ecuaciones diferenciales parciales denominadas ecuaciones de difusión. Las ecuaciones de difusión permiten describir el comportamiento de los alelos mutantes teniendo en cuenta los cambios aleatorios resultantes de la elección al azar de los gametos (células germinales) en la reproducción, así como los cambios deterministas producidos por la mutación y la selección. Aunque el método de las ecuaciones de difusión comporta cierta inexactitud, da respuesta a cuestiones importantes y difíciles que por otros métodos resultarían inaccesibles, tales como: ¿cuál es la probabilidad de fijación de un mutante que aparece en una población finita y muestra cierta ventaja selectiva? Es decir, ¿cuál es la probabilidad de que ese gen se propague finalmente por toda la población?

Sin embargo, la aplicabilidad de este método a los cambios genéticos de la evolución permaneció bastante limitada durante algún tiempo. La razón de ello es que la genética de poblaciones trabaja con el concepto de frecuencias génicas (el predominio relativo de varios alelos en una población), mientras que los estudios evolutivos convencionales se realizaban a nivel fenotípico, no habiendo por aquel entonces forma directa de conectar ambos grupos de datos sin caer en la ambigüedad. Este obstáculo se salvó con el desarrollo de la genética molecular. Hizo posible comparar en individuos genéticamente próximos, moléculas específicas de ARN (el producto directo de los genes) y proteínas (el producto final) y, por tanto, determinar la velocidad a la que los genes alélicos son sustituidos durante la evolución. Permitió, asimismo, el estudio de la variabilidad de los genes dentro de una misma especie. Por fin había llegado el momento de aplicar la teoría matemática de la genética de poblaciones al objeto de descubrir cómo evolucionaban los



aminoácidos que constituye la cadena. La molécula de hemoglobina posee dos cadenas alfa y dos beta, que se originaron mediante la duplicación de un único gen hace unos 450 millones de años. La tabla refleja la casi uniformidad (predicha por la teoría neutralista, que defiende el autor) de la tasa evolutiva de una determinada proteína en organismos muy diferentes. El número de diferencias en los aminoácidos es aproximadamente 20 cuando se comparan entre sí cualquiera de los tres mamíferos y aproximadamente 70 si se compara la carpa con cualquiera de los tres mamíferos.

TIPO DE CAMBIO	ALFA HUMANA RESPECTO BETA HUMANA	ALFA DE LA CARPA RESPECTO BETA HUMANA
SIN CAMBIO	62	61
UN NUCLEOTIDO	55	49
DOS NUCLEOTIDOS	21	29
ADICION O DELECION	9	10
TOTAL	147	149

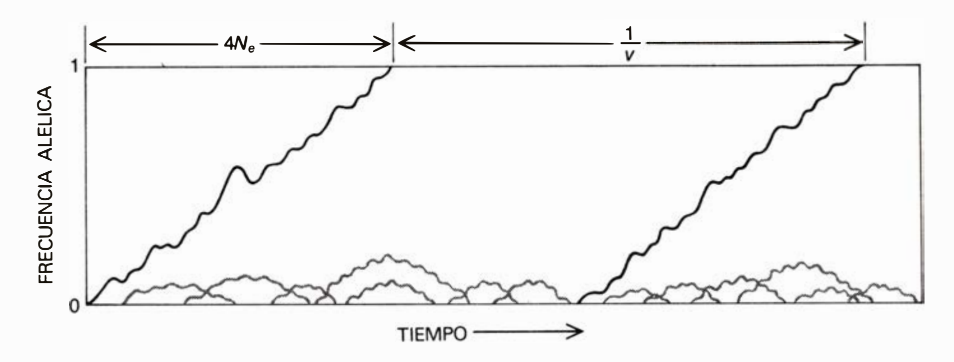
EL NUMERO DE DIFERENCIAS entre las secuencias de aminoácidos de la cadena alfa y la cadena beta de la hemoglobina humana, comparado con el número de diferencias entre las secuencias de la cadena alfa de la carpa y la cadena beta humana. La columna de la izquierda clasifica los restos aminoácidos: si no hay cambio, si existe un cambio debido como mínimo a una sustitución nucleotídica o como mínimo a dos sustituciones en el código genético de cada resto, o si se da una adición o deleción de un aminoácido. Los valores son similares tanto si se comparan las cadenas de una misma especie como de las dos especies, lo que sugiere que las cadenas alfa han acumulado mutaciones más o menos en la misma proporción en las dos líneas filogenéticas a lo largo de 400 millones de años.

genes. Cabía esperar que el principio de la selección darwinista prevaleciera a este nivel fundamental. Y así, muchos biólogos evolutivos hallaron lo que esperaban encontrar, y propendieron a extender el panseleccionismo hasta el nivel molecular.

Teoría neutralista

El panorama de cambios evolutivos que en realidad surge de los estudios moleculares me pareció, sin embargo, incompatible con los resultados que cabía esperar desde un punto de vista neodarwinista. Uno de mis hallazgos más destacados fue que, para una proteína determinada, la tasa de sustitución de un aminoácido (la subunidad de las proteínas) por otro es aproximadamente igual en muchas líneas filogenéticas distintas. Otro descubrimiento fue que estas sustituciones, en vez de seguir un modelo, parecían ocurrir al azar. El tercer descubrimiento hacía referencia a que la tasa total de cambio a nivel de ADN, el material genético, era muy alta: del orden de la sustitución de por lo menos una base nucleotídica (subunidad del ADN) por genoma (dotación genética total) cada dos años en una línea evolutiva de mamíferos. En cuanto a la variabilidad dentro de una especie, los métodos electroforéticos de detección de pequeñas diferencias entre proteínas revelaron inesperadamente una gran cantidad de variabilidad genética; en diversos organismos, se vio que las proteínas producidas por una gran parte de los genes eran polimórficas, es decir, estaban presentes en la especie en varias formas. En muchos casos, los polimorfismos proteínicos no tenían efectos fenotípicos visibles ni una correlación obvia con las condiciones ambientales.

En 1967, al considerar estas problemáticas observaciones, llegué a la conclusión de que sugerían dos cosas. Por un lado, la mayoría de las sustituciones de nucleótidos acaecidas en el transcurso de la evolución debían ser el resultado de la fijación al azar de mutantes neutros, o casi neutros, más que el resultado de una selección darwinista positiva. Por otro, muchos de los polimorfismos proteínicos debían ser selectivamente neutros o casi neutros y su persistencia en una población se debería al equilibrio existente entre la aportación de polimorfismo por mutación y su eliminación al azar. Expuse estas ideas en una conferencia del Genetics Club en Fukuoka en noviembre de 1967, y en un breve artículo que apareció en *Nature* en febrero del año siguiente. En 1969 obtuve un gran apoyo en un artículo de Jack Lester King, ahora en la Universidad de California en Santa Bárbara, y Thomas H. Jukes, de la Universidad de California en Berkeley, que publicaron en *Science*. Habían llegado independientemente a conclusiones similares acerca de la evolución molecular (aunque no sobre los polimorfismos proteínicos), y aportaban datos convincentes procedentes de la biología molecular.



ALELOS MUTANTES (genes variantes). Aparecen al azar en la población. Su frecuencia fluctúa; la mayoría de ellos desaparecen con el tiempo (*gris*), pero algunos se difunden por la población hasta llegar a la fijación: una frecuencia igual a la unidad, o al 100 por cien (*negro*). Los estudios de genética de poblaciones revelan que un alelo neutro destinado a la fijación necesita un número promedio de generaciones igual a cuatro veces el tamaño eficaz de la población, o 4 N_e, para alcanzarla. La cifra de generaciones entre dos fijaciones consecutivas es igual al recíproco de la tasa de mutación v.

Los artículos que apoyaban una teoría neutralista fueron duramente criticados por los evolucionistas, que afirmaban que los nuevos datos moleculares podían interpretarse a la luz de los principios ortodoxos neodarwinistas. La controversia neutralismo-seleccionismo continúa todavía en la actualidad. La diferencia esencial entre las dos escuelas teóricas puede apreciarse comparando las respectivas interpretaciones del proceso evolutivo por medio del cual se sustituyen los genes mutantes en una especie. Cada sustitución comprende una serie de pasos en los que aparece un alelo raro en una población y finalmente se difunde por toda ella hasta alcanzar la fijación, es decir, una frecuencia del 100 por cien. Los seleccionistas sostienen que, para que un alelo mutante se difunda en una especie, debe poseer alguna ventaja selectiva (aunque admiten que un alelo, en sí neutro, puede ocasionalmente ser arrastrado junto a un gen que se está seleccionando y con el que se halla fuertemente unido, pudiendo alcanzar así una frecuencia alta).

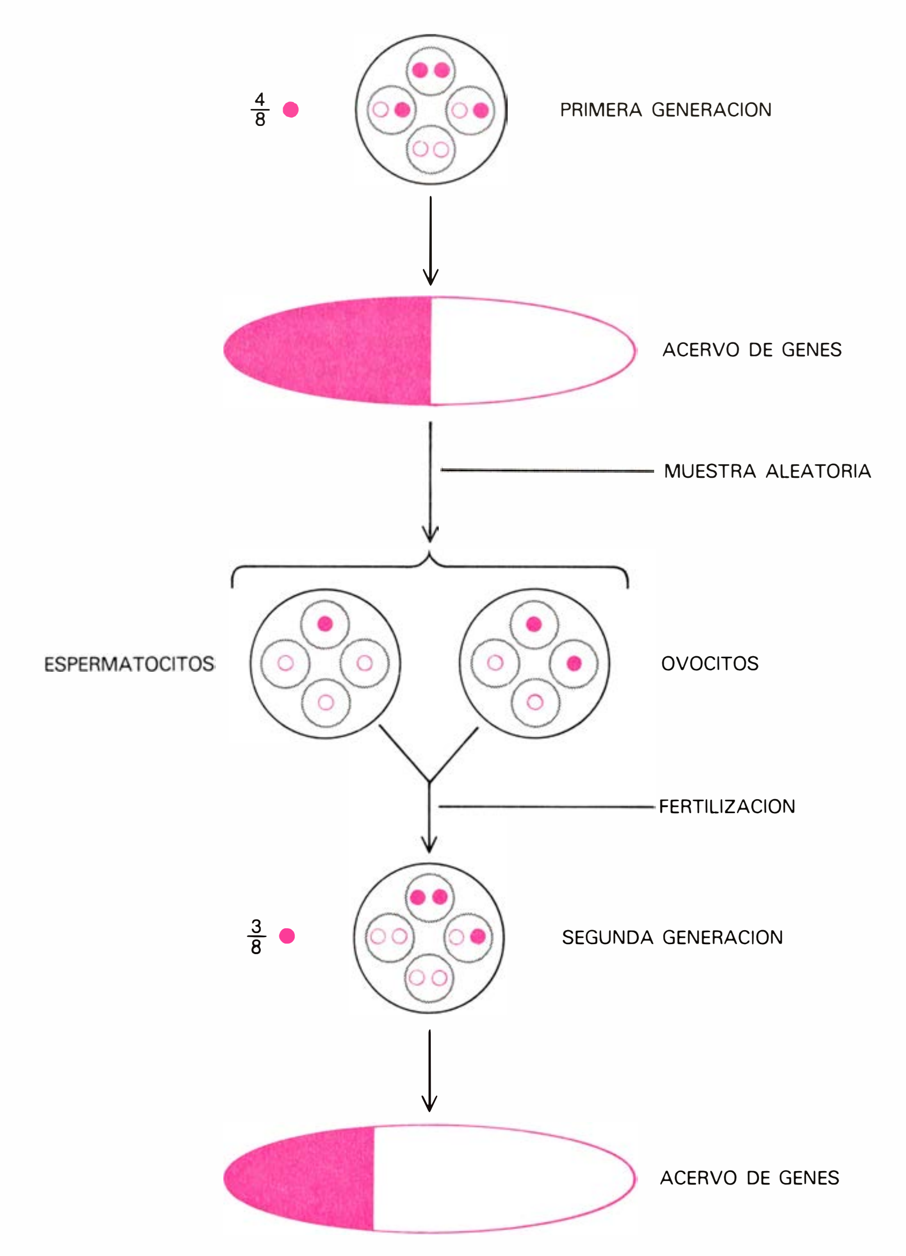
Por otra parte, los neutralistas defienden que algunos mutantes pueden difundirse en una población sin tener ninguna ventaja selectiva. Si un mutante es selectivamente equivalente a los alelos preexistentes, su suerte depende del azar. Su frecuencia fluctúa, incrementándose o decreciendo fortuitamente con el tiempo, porque sólo se escoge un número relativamente pequeño de gametos, de entre el amplio número de gametos masculinos y femeninos producidos en cada generación, y están, por tanto,

representados en los individuos de la generación siguiente [*véase la ilustración de la derecha*].

En el curso de esta deriva aleatoria, la inmensa mayoría de los alelos mutantes se pierden por azar, pero la fracción restante termina por fijarse en la población. Si las mutaciones neutras son frecuentes a nivel molecular y si la deriva genética es continua durante un largo período de tiempo (digamos millones de generaciones), la composición genética de la población cambiará significativamente. Para cualquier mutante neutro que aparece en una población, la probabilidad de fijación es igual a su frecuencia inicial. El tiempo promedio necesario hasta la fijación (excluyendo los alelos que se pierden) es cuatro veces el tamaño "eficaz" de la población, o $4N_e$. (El tamaño eficaz de una población es aproximadamente igual al número de individuos reproductores de una generación, y suele ser mucho menor que el número total de individuos de una especie.)

Debo aclarar que la teoría neutralista no supone que los genes neutros no son funcionales, sino sólo que varios alelos pueden ser igualmente eficaces para promover la supervivencia y reproducción del individuo. Si un alelo mutante contiene aminoácidos diferentes en una proteína, la proteína modificada no necesita más que funcionar tan bien como la forma original, pero no tiene por qué ser exactamente equivalente. En particular, en los organismos superiores, la homeostasis contrarresta los cambios ambientales al igual que los cambios fisiológicos internos; las fluctuaciones del ambiente no implican necesariamente que debe haber fluctuaciones comparables en la eficacia darwiniana de los genes mutantes.

Algunas críticas a la teoría neutralista proceden de una incorrecta definición de la "selección natural". Esta expresión debería aplicarse estrictamente en su sentido darwiniano: la selección natural actúa a través de —y debe venir determinada por— las diferencias en la supervivencia y reproducción del individuo. La mera existencia de diferencias funcionales detectables entre dos formas moleculares no prueba la actuación de la selección natural, la cual sólo puede determinarse mediante la investigación de las tasas de superviviencia y fecundidad. Además debería hacerse una clara distinción entre selección positiva (darwiniana) y negativa. Esta última, de la que Muller demostró que constituía la forma más común, elimina los mutantes deletéreos y tiene poco que ver con las sustituciones génicas de la evolución. El descubrimiento de la selección negativa



CAMBIOS AL AZAR de las frecuencias génicas. Dichos cambios proceden de un muestreo aleatorio de los gametos (células germinales) en la reproducción, como se muestra aquí en una población hipotética de cuatro individuos (*círculos grises*), cada uno de los cuales posee dos genes homólogos (*círculos y circunferencias de color*), heredados del padre y de la madre. En la primera generación, la frecuencia del alelo "de color" es de 4/8 y, por tanto, representa el 50 por ciento del acervo genético. De los muchos gametos que se producen en una generación, sólo unos pocos son escogidos, al azar, en la reproducción. Aquí resulta que sólo un alelo de color está presente en los cuatro gametos masculinos de la primera generación comprometidos en la reproducción, de modo que la frecuencia del alelo en color pasa a ser de 3/8 en la segunda generación de individuos y, por tanto, en su acervo o "pool" de genes.

no contradice la teoría neutralista. Finalmente, debe recordarse la distinción entre mutación génica en el individuo y sustitución génica en la población; sólo la última está directamente relacionada con la evolución molecular. Para los mutantes ventajosos la tasa de sustitución está altamente condicionada por el tamaño de la población y por el grado de ventaja selectiva (como demostraré más adelante) así como por la tasa de mutación.

Dos importantes descubrimientos relativos a la evolución molecular demuestran con especial claridad que sus modelos son completamente diferentes de los de la evolución fenotípica y que las leyes que rigen las dos formas de evolución son distintas. El primer descubrimiento, aludido anteriormente, reveló que, para cada proteína, la tasa de evolución en términos de sustituciones de aminoácidos por año era prácticamente constante y muy parecida en distintas líneas filogenéticas. El otro ponía de manifiesto que las moléculas, o partes de una molécula, sometidas a un grado de limitación funcional relativamente pequeño tenían una tasa de evolución más alta (en términos de sustituciones mu-

tantes) que las que se hallan sometidas a limitaciones mayores.

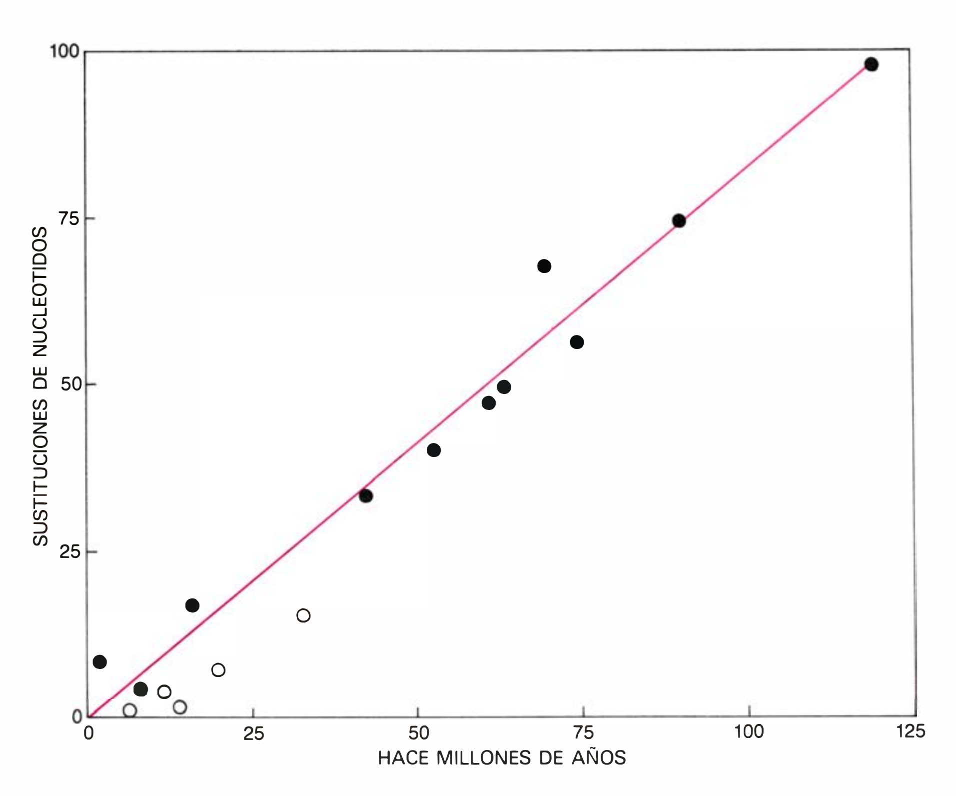
Evolución molecular

La constancia de la tasa evolutiva se hace presente en la molécula de hemoglobina, que en los peces óseos y en los vertebrados superiores es un tetrámero (una molécula con cuatro subunidades grandes) formado por dos cadenas alfa idénticas y dos cadenas beta idénticas. En los mamíferos, los aminoácidos se sutituyen en la cadena alfa, que tiene 141 aminoácidos, en una proporción de aproximadamente un cambio cada siete millones de años. Esto corresponde más o menos a una sustitución cada 1000 millones de años (10^{-9} sustituciones al año) por resto aminoácido. Esta tasa no parece depender de factores tales como el tiempo de generación, las condiciones de vida o el tamaño de la población. El valor prácticamente constante de la tasa evolutiva se pone de manifiesto cuando el número de aminoácidos diferentes entre las cadenas alfa de varios vertebrados se compara con el árbol filogenético que muestra las relaciones entre los vertebrados y sus tiempos de divergencia evolutiva [*véanse las ilustraciones de las páginas 46 y 47*].

Las cadenas alfa y beta tienen esencialmente la misma estructura, más o menos la misma longitud y presentan aproximadamente la misma tasa evolutiva de sustitución de aminoácidos. Aparecieron a través de una duplicación génica hace 450 millones de años y se diferenciaron al acumular mutaciones independientemente. Si se compara la divergencia entre la cadena alfa y la beta del hombre con la divergencia entre la cadena alfa de la carpa y la beta del hombre, se hace evidente que en ambos casos las cadenas alfa y beta difieren entre sí aproximadamente lo mismo. El hecho de que la cadena alfa del hombre y la de la carpa difieran aproximadamente en la mitad de sus restos aminoácidos, sugiere que las cadenas alfa de las dos líneas filogenéticas distintas, una que conduce a la carpa y otra que conduce al hombre, han acumulado mutaciones independientemente y casi en la misma proporción durante un período de tiempo de unos 400 millones de años. Es más, la tasa de sustitución de aminoácidos que se observa en estas comparaciones es muy similar a las observadas al comparar las cadenas alfa de varios mamíferos.

Mi tesis sobre la constancia de la tasa evolutiva a nivel molecular ha sido criticada, entre otros, por Richard C. Lewontin, de la Universidad de Harvard, que dice refiriéndose a dicha constancia: "una simple confusión entre un promedio y una constante" y "nada más que la ley de los grandes números". Estas observaciones revelan una concepción errónea de la naturaleza de la evolución molecular. Se intenta comparar aquí las tasas evolutivas intrínsecas de diferentes líneas filogenéticas. Las tasas de mortalidad características del hombre y de un insecto no llegan a ser iguales por el mero hecho de que se tome el valor promedio de un largo período de tiempo o de un gran número de individuos; no existe ninguna razón para esperar que dos promedios converjan a menos que los factores intrínsecos que los generan sean iguales. Mi opinión es que las tasas evolutivas intrínsecas están determinadas esencialmente por la estructura y función de las moléculas y no por las condiciones ambientales.

Las tasas evolutivas no son exactamente constantes en el sentido en que lo es una tasa de desintegración radiactiva. Mi colega Tomoko Ohta y yo demostramos en 1971 que la varianza (el cuadrado de la desviación típica) de la tasa evolutiva observada para las hemoglobinas y para el citocromo *c* de diferentes líneas de mamíferos es de 1,5 a 2,5 veces mayor que la varianza esperada si se debiera solamente al azar. Charles H. Langley y Walter M. Fitch realizaron un análisis más elaborado en la Facultad de Medicina de la Universidad de Wisconsin, combinando los datos de las cadenas alfa y beta de hemoglobina, el citocromo *c* y el fibrinopéptido *A*. Encontraron que la varianza de las tasas de sustituciones mutantes era unas 2,5 veces mayor que la esperada debida a las fluctuaciones al azar, y tomaron este hecho como una prueba contra la teoría neutralista. También demostraron que cuando se representa el número estimado de sustituciones entre ramas divergentes de un árbol filogenético frente al correspondiente tiempo de divergencia, los puntos forman una línea recta, lo que indica la sustancial uniformidad de las tasas evolutivas. Me parece incorrecto sobreestimar las fluctuaciones locales y elevarlas a la categoría de prueba contra la teoría neutralista y no querer investigar por qué la tasa permanece esencialmente constante.



SUSTITUCIONES DE NUCLEOTIDOS, estimadas a partir del número total de diferencias en los aminoácidos, observadas en siete proteínas de 16 pares de mamíferos. Se representan en función del tiempo de divergencia de los miembros de cada par. Excepto para las líneas filogenéticas de los primates (*círculos en blanco*), los puntos forman casi una línea recta; ello sugiere de nuevo la uniformidad de la tasa evolutiva molecular de una proteína. Datos recogidos por Walter M. Fitch, de la Facultad de Medicina de Wisconsin. En abcisas, los números corresponden a millones de años.

Tasas evolutivas

Volviendo a las relaciones cuantitativas que determinan las tasas evolutivas, consideremos primero los nucleótidos

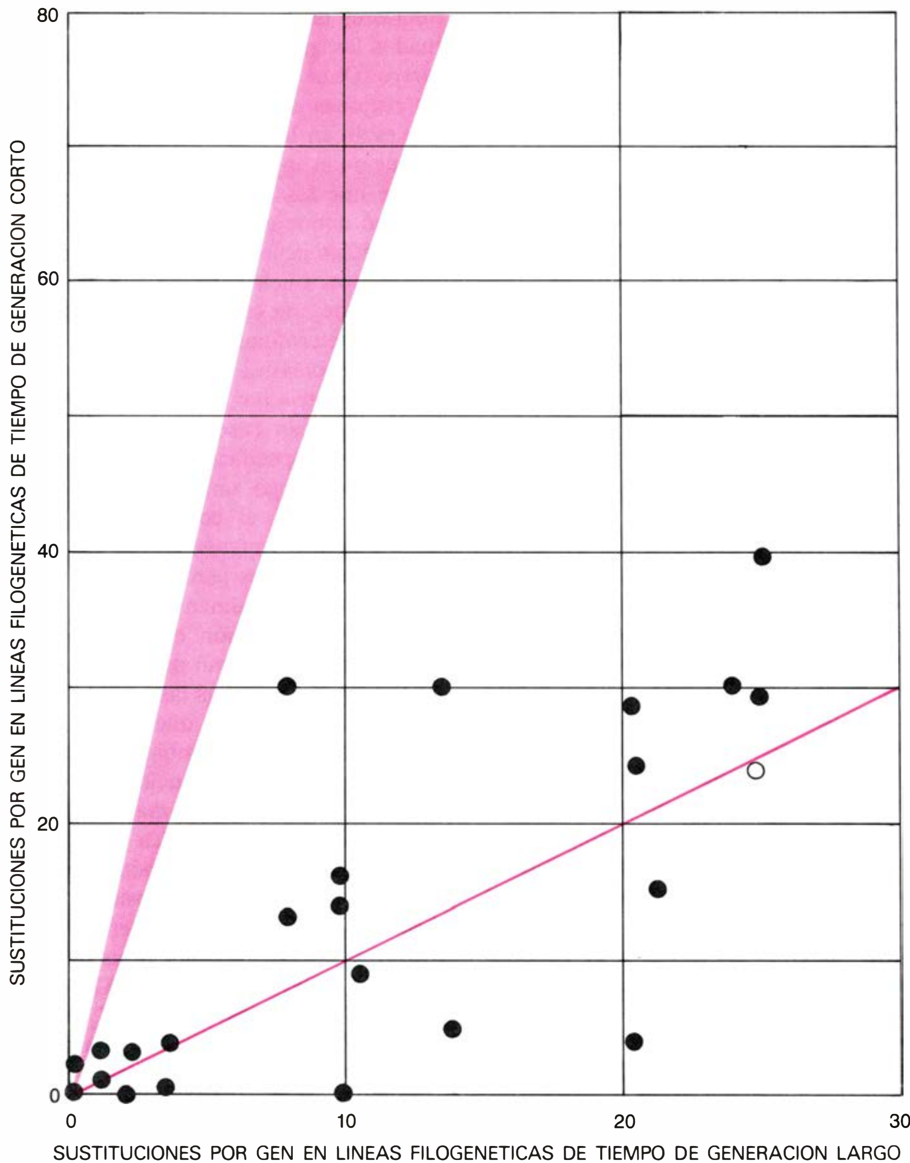
que constituyen un genoma: una única (haploide) dotación de cromosomas. En un ser humano, el número de nucleótidos es muy grande, del orden de 3500 millones. Debido a que la tasa de mutación por unidad nucleotídica es baja (quizás 10^{-8} por generación, o una mutación por cada 100 millones de generaciones) se puede suponer que siempre que aparece un mutante lo hace en un nuevo resto o unidad. Esta suposición se llama en genética de poblaciones el modelo de "los infinitos restos".

Sea ***n*** la tasa de mutación por gameto y por unidad de tiempo (generación). Puesto que cada individuo posee dos complementos cromosómicos, el número total de mutantes nuevos introducidos, en cada generación, en una población de N individuos es $2\ N\nu$. Sea ahora u la probabilidad de que un mutante logre, por fin, la fijación. Entonces, en un estado de equilibrio en el que el proceso de sustitución continúa durante largo tiempo, la tasa k de sustituciones mutantes por unidad de tiempo viene dada por la ecuación $k = 2\ N\nu u$. Es decir, aparecen $2\ N\nu$ nuevos mutantes en cada generación, de los que la fracción u logra fijarse, y k representa la tasa evolutiva en función de las sustituciones mutantes. Esta ecuación puede aplicarse no sólo a todo el genoma sino también, y con bastante exactitud, a un único gen formado por varios centenares de nucleótidos o a la proteína codificada por un gen.

La probabilidad u de fijación es una cantidad bien conocida en genética de poblaciones. Si el mutante es selectivamente neutro, u es igual a $1/(2N)$. La razón de ello es que cualquiera de los $2N$ genes de la población tiene la misma probabilidad de fijarse que los demás; por tanto, la probabilidad de que el nuevo mutante sea el gen afortunado será de $1/(2N)$. (Esto supone que se considera el proceso durante un largo período de tiempo, ya que el tiempo promedio para que un gen neutro se extienda en la población es $4\ N_e$.) Sustituyendo $1/(2N)$ por u en la ecuación de la tasa evolutiva ($k = 2\ N\nu u$), se obtiene $k = \nu$. Es decir, la tasa evolutiva en función de las sustituciones mutantes en la población equivale simplemente a la tasa de mutación por gameto, con independencia de cuál sea el tamaño de la población.

Esta importante relación rige sólo para los alelos neutros. Si el mutante tiene una pequeña ventaja selectiva s, entonces u es aproximadamente igual a $2s$ y la ecuación de la tasa evolutiva se convierte en $k = 4\ Ns\nu$. Es decir, la tasa evolutiva para los genes con ventaja selectiva depende del tamaño de la población, de la ventaja selectiva y de la proporción en que los mutantes con una ventaja selectiva determinada aparecen en cada generación. En este caso podría esperarse que la tasa evolutiva dependiera mucho del ambiente, siendo alta para una especie a la que se le ofrecen nuevas oportunidades ecológicas pero baja para las que se mantienen en un ambiente estable. Es muy improbable, creo yo, que el producto $Ns\nu$ sea el mismo para las diversas líneas filogenéticas de vertebrados, en algunas de las cuales la evolución fenotípica ha sido muy rápida (como en la línea que conduce al hombre), mientras que en otras la evolución fenotípica hace tiempo que prácticamente ha cesado (como en la línea que conduce a la carpa). Aún así las tasas de evolución molecular observadas presentan una notable constancia. Me parece que esta constancia es mucho más compatible con las previsiones de la teoría neutralista, es decir, con la ecuación $k = \nu$ que con la relación seleccionista $k = 4\ Ns\nu$.

Más sorprendente aún que la constancia de la tasa evolutiva es la segunda



TASA DE EVOLUCION MOLECULAR de los mamíferos que tienen un tiempo de generación corto en comparación con la de los mamíferos que tienen un tiempo de generación largo. El diagrama se basa en un estudio de Allan C. Wilson y sus colegas de la Universidad de California en Berkeley. Cada punto representa la proporción de sustituciones de nucleótidos ocurridas en los dos animales de un par desde que divergieron de un antecesor común; el círculo en blanco, por ejemplo, representa la cadena beta de hemoglobina en el elefante (*abcisas*) y en el ratón (*ordenadas*). Si la tasa de cambio por año fuera idéntica para los dos animales de un par, los puntos formarían una línea recta (*línea de color*). En realidad, los puntos caen cerca de esta línea de tiempo absoluto y lejos del sector predicho para el efecto del tiempo de generación (*área coloreada*). La tasa de evolución molecular es casi constante por año.

característica fundamental de la evolución molecular: cuanto menor es la limitación funcional en una molécula o en una parte de una molécula, mayor es la tasa evolutiva de sustituciones mutantes. Por ejemplo, hay regiones de ADN entre genes, y en el caso de organismos superiores incluso dentro de los genes, que no participan en la formación de la proteína y, por tanto, deben estar mucho menos sujetas a la selección natural; algunas investigaciones recientes han puesto de manifiesto que las sustituciones nucleotídicas son especialmente frecuentes en tales regiones del ADN.

Limitación funcional

Esta correlación entre una relativa falta de presión selectiva y una tasa de evolución molecular relativamente alta ha podido ser bien estudiada en ciertas proteínas. Entre las proteínas investigadas hasta ahora, la tasa evolutiva más alta se ha encontrado en los fibrinopéptidos, que parecen tener una escasa función, si es que tienen alguna, después de separarse del fibrinógeno para producir la fibrina, proteína que desempeña un papel importante en la coagulación de la sangre. El mismo efecto se observa en el caso de la cadena C de la molécula de proinsulina, un precursor de la insulina. La cadena C, que se separa del precursor para formar la insulina activa, evoluciona a una tasa varias veces mayor que la molécula activa. El efecto de la limitación funcional sobre la tasa evolutiva también se ha hecho notar en diferentes partes de la molécula de hemoglobina. La estructura de la superficie de la molécula es seguramente menos importante que la estructura de las cavidades interiores de la misma, donde están situados los grupos hemo, que contienen hierro. Ohta y yo hemos calculado que las regiones de las cadenas alfa y beta que están en la superficie de la proteína evolucionan unas 10 veces más rápidamente que las regiones que forman la cavidad donde está el grupo hemo.

El código genético está basado en grupos de tres nucleótidos, y cada triplete "codón" de una cadena de ARN especifica un determinado aminoácido de la cadena proteínica codificada por el ARN. Por ejemplo, el codón *GUU* (las letras significan bases nucleotídicas determinadas) especifica al aminoácido valina. Sin embargo, también lo especifica el codón *GUC*; el código genético es "degenerado", designándose la mayoría de los aminoácidos por dos o más sinónimos, que normalmente difieren sólo en la tercera posición del triplete. En consecuencia, una gran parte (quizás un 70 por ciento) de todas las sustituciones de nucleótidos al azar que afectan a la tercera posición son cambios sinónimos que no conducen a sustituciones de aminoácidos. Cada día se ve más claro que la sustitución nucleotídica evolutiva tiene lugar en una proporción especialmente alta en la tercera posición. Michael Grunstein, de la Universidad de California en Los Angeles, y sus colegas compararon las secuencias de ARN que codifican la proteína histona IV en dos especies de erizos de mar. Encontraron que aunque la proteína ha mantenido una secuencia de aminoácidos prácticamente invariable durante aproximadamente 1000 millones de años, se encuentran muchas diferencias de nucleótidos sinónimos en las secuencias de ARN de las dos especies. Basándome en sus datos y en pruebas paleontológicas del tiempo de divergencia de estas especies, he estimado que la tasa de sustitución nucleotídica en la tercera posición ha sido aproximadamente de $3,7 \times 10^{-9}$ por año, una tasa verdaderamente alta. Lo extraordinario es que haya habido tantas sustituciones mutantes sinónimas en el gen de la histona IV a pesar de la baja tasa de sustituciones de aminoácidos en la proteína correspondiente.

Estas observaciones pueden explicarse de manera simple y coherente por medio de la teoría neutralista. Supongamos que una cierta fracción f_0 de los mutantes moleculares sean selectivamente neutros y que los restantes sean totalmente deletéreos. Entonces, la tasa de mutación ν para los alelos neutros es igual a la tasa de mutación total ν_T multiplicada por f_0, de modo que la tasa global de sustituciones mutantes k sería igual a $\nu_T f_0$. Supongamos ahora que la probabilidad de que un cambio mutacional sea neutro (no perjudicial) dependa en gran manera de la limitación funcional. Cuanto menor sea la limitación, mayor será la probabilidad f_0 de que un cambio al azar sea neutro, aumentando en consecuencia la tasa evolutiva k. La tasa evolutiva máxima se logra cuando f_0 es igual a 1, es decir, cuando todas las mutaciones son neutras. En mi opinión, las altas tasas evolutivas observadas en la tercera posición del codón están muy cerca de este límite.

La teoría neutralista predice, por tanto, que cuando disminuye la limitación funcional, la tasa evolutiva converge hacia el valor máximo establecido por la tasa de mutación total. La confirmación, por estudios posteriores, de dicha convergencia, o uniformización, de las tasas de evolución molecular representaría un gran apoyo a la teoría neutralista. Esta interpretación de los datos moleculares no tendrá ningún sentido para los seleccionistas. En su opinión, las moléculas o partes de una molécula que evolucionan rápidamente en términos de sustituciones mutantes deben encerrar alguna función importante, aunque desconocida hasta ahora, y deben experimentar una rápida mejora adaptativa por medio de la acumulación de mutantes beneficiosos. Y no verán razón alguna para creer que el límite superior de la tasa evolutiva esté relacionado con la tasa de mutación total.

Polimorfismo

Los neutralistas y seleccionistas han dado también explicaciones diametralmente opuestas a los mecanismos me-

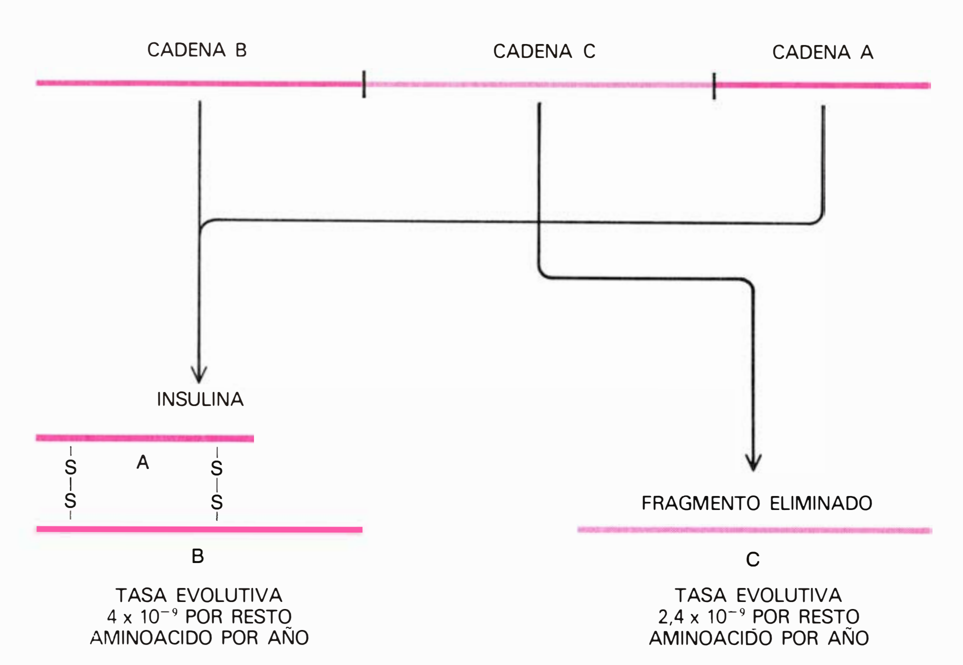
PROTEINA	TASA EVOLUTIVA
FIBRINOPEPTIDOS	9,0
RIBONUCLEASA PANCREATICA	3,3
CADENAS DE HEMOGLOBINA	1,4
MIOGLOBINA	1,3
LISOZIMA ANIMAL	1,0
INSULINA	0,4
CITOCROMO *c*	0,3
HISTONA IV	0,006

LAS PROTEINAS SE DIFERENCIAN AMPLIAMENTE (*izquierda*) en su tasas evolutiva (*derecha*). En el esquema se dan las tasas evolutivas de varias proteínas en función del número de sustituciones de aminoácidos por resto aminoácido y por 1000 millones de años. Dicha tasa es especialmente alta para los fibrinopéptidos, que parecen tener una función poco importante, una vez separados de una molécula precursora para producir la fibrina, proteína activa de la coagulación de la sangre. Proteínas como los fibrinopéptidos están sometidos a menos limitaciones funcionales, y evolucionan más deprisa, que las proteínas cuya estructura exacta resulta ser importante y, por tanto, sufren una mayor limitación.

diante los cuales se mantiene la variabilidad genética en una especie, en particular en forma de polimorfismo proteínico: la coexistencia en una especie de dos o más formas diferentes de una proteína. Los neutralistas defienden que el polimorfismo es selectivamente neutro y que se mantiene en una población mediante el aporte mutacional y la eliminación al azar; en cada generación aparece cierto número de mutantes neutros que, con el tiempo, o se fijan en la población, o se pierden; durante ese proceso contribuyen a la variabilidad genética a través del polimorfismo. Desde el punto de vista neutralista, el polimorfismo y la evolución molecular no son dos fenómenos distintos; el polimorfismo es sólo una fase de la evolución molecular.

Los seleccionistas sostienen que los polimorfismos se mantienen activamente gracias a alguna forma de "selección equilibradora"; y destacan, a este respecto, la selección heterótica, o "ventaja de los heterocigotos", y la selección dependiente de las frecuencias. Durante algún tiempo, la selección heterótica fue defendida con entusiasmo como el principal agente responsable del mantenimiento de los polimorfismos. Hay ejemplos en que los individuos heterocigotos para un gen determinado (que poseen un alelo diferente del gen en cada uno de sus dos cromosomas) son más aptos que los individuos homocigotos para cualquiera de los dos alelos (poseen uno u otro alelo en ambos cromosomas). La selección tendería entonces a conservar ambos alelos en la población como un polimorfismo equilibrado. Sin embargo, en 1973, Roger D. Milkman, de la Universidad de Iowa, encontró abundante polimorfismo en la bacteria *Escherichia coli*, que es un organismo haploide: posee solamente una dotación de genes. La ventaja de los heterocigotos no puede explicar tales polimorfismos.

Hoy en día muchos seleccionistas explican el polimorfismo como resultado de una selección dependiente de la frecuencia, en la que la eficacia de dos alelos varía con su frecuencia relativa. Esta teoría fue propuesta por primera vez por Ken-Ichi Kojima y sus colegas, de la Universidad de Texas en Austin; obtuvieron resultados que indicaban que una selección dependiente de la frecuencia afectaba visiblemente a los genes que codificaban los enzimas esterasa-6 y alcohol deshidrogenasa (ADH) de la mosca del vinagre *Drosophila melanogaster*. Bryan Clarke, de la Universidad de Nottingham, comunicó que había confirmado los resultados de Kojima en el caso de la ADH. Por otra parte, los experimentos realizados por Tsuneyuki Yamazaki, ahora en la Universidad de Kyushu, no pudieron demostrar ninguna selección de este tipo para los alelos de la esterasa-5 de *Drosophila pseudoobscura*. Un grupo de investigadores dirigidos por Terumi Mukai, de Kyushu, llevaron a cabo estudios muy amplios de selección para varios enzimas de *D. melanogaster* y no encontraron ningún indicio de que hubieran diferencias de eficacia entre las diversas formas de los enzimas. Tampoco los experimentos recientes, a gran escala, realizados por Mukai e Hiroshi Yoshimaru han logrado encontrar ninguna selección dependiente de la frecuencia para la ADH de *D. melanogaster*.



CADENA C DE LA PROINSULINA (*color claro*), **cortada de la molécula precursora y desechada. Las cadenas A y B** (*color oscuro*) **se unen por medio de puentes disulfuro** (*S-S*) **para formar la molécula activa de insulina. Dada la relativa libertad de la cadena C en lo que respecta a las limitaciones funcionales, evoluciona de un modo mucho más rápido que las dos cadenas de la molécula activa.**

Si la selección no es la responsable del mantenimiento de los polimorfismos, ¿qué explicación neutralista existe para el hecho de que algunas proteínas sean más polimórficas que otras? Richard K. Koehn, de la Universidad de Nueva York en Stony Brook, y W. F. Eanes, ahora en Harvard, y también el grupo de Mastoshi Nei, de la Universidad de Texas en Houston, han demostrado que en varias especies de *Drosophila* existe una correlación significativa entre la variabilidad genética (o polimorfismo) de las proteínas y el peso de sus subunidades moleculares. Esto halla fácil justificación en la teoría neutralista: cuanto mayor sea el tamaño de una subunidad, más alta será su tasa de mutación. Harry Harris, de la Universidad de Pennsylvania, y sus colegas no pudieron encontrar la misma correlación cuando investigaron los polimorfismos humanos, pero observaron que los enzimas que tienen una única subunidad son más polimórficos que los que poseen múltiples subunidades, lo que ya había sido demostrado en *Drosophila* por Eleftherios Zouros, de la Universidad de Dalhousie. Uno de los descubrimientos de Zouros y Harris se ajusta especialmente bien a la teoría neutralista: los enzimas con múltiples subunidades que forman moléculas híbridas al combinarse con enzimas codificados por otros genes ofrecen, sin la menor ambigüedad, un reducido nivel de polimorfismo. La precisa interacción entre subunidades requerida para formar tales enzimas incrementaría el grado de limitación funcional y reduciría, por tanto, la posibilidad de que una mutación fuera inocua, o neutra.

En otras palabras, los neutralistas consideran que las principales causas determinantes del polimorfismo proteínico son la estructura y la función molecular. Para los seleccionistas, las principales causas determinantes son las condiciones ambientales. Estos últimos han sostenido que debía existir una correlación entre la variabilidad ambiental y la genética. Predijeron, por ejemplo, que los organismos que vivían en el fondo del mar manifestarían, en general, poca variabilidad genética porque su ambiente era estable y homogéneo, en tanto que los organismos que vivían en la zona intermareal manifestarían una gran variabilidad genética porque su ambiente era

ARN MENSAJERO DEL *L. PICTUS*	GA U	AAC	AUC	CAA	GG A	AU· A	AC U	AAA	CCG	GC A	AUC
ARN MENSAJERO DEL *S. PURPURATUS*	GA C	AAC	AUC	CAA	GG U	AU C	AC G	?	?	GC U	AUC
SECUENCIA DE AMINOACIDOS DE LA HISTONA IV EN AMBAS ESPECIES	Asp	Asn	Ile	Gln	Gly	Ile	Thr	Lys	Pro	Ala	Ile
	24	25	26	27	28	29	30	31	32	33	34

SECUENCIA DE NUCLEOTIDOS del ARN mensajero que codifica la proteína histona IV de dos especies de erizos de mar. *Lytechinus pictus* y *Strongylocentrotus purpuratus*, en una comparación realizada por Michael Grunstein, de la Universidad de California en los Angeles. Existen cuatro nucleótidos en el ARN (*A, G, U,* y *C*); los codones, formados por tres nucleótidos, especifican los diversos aminoácidos que constituyen una proteína. La mayoría de los aminoácidos son especificados por dos o más codones sinónimos, que normalmente sólo difieren en la tercera posición. En esta corta porción de ARN que codifica desde el resto aminoácido 24 hasta el 34 de la histona IV, existen cinco diferencias sinónimas (*en color*) en los nucleótidos de la tercera posición. Ha ocurrido una alta tasa de sustitución de nucleótidos en la posición tres, que apenas sufre limitaciones.

cambiante. Esta predicción lógica y plausible, no tuvo éxito: se ha descubierto que la variabilidad genética suele ser muy alta entre los organismos que viven en el fondo de los océanos y, muy baja, entre los que viven en la zona intermareal.

Modelos

Para acometer estudios cuantitativos basados en la genética de poblaciones se necesitan modelos matemáticos que den cuenta de la producción mutacional de nuevos alelos. El primero de estos modelos fue propuesto en 1964 por James F. Crow, de la Universidad de Wisconsin, y por el autor de este artículo. Se basa en el hecho de que cada gen está formado por un gran número de nucleótidos, de modo que puede aparecer un número prácticamente infinito de alelos. El modelo supone, pues, que cualquier nuevo mutante que surja representa un alelo nuevo y no uno ya preexistente. El modelo predice que la variabilidad dentro de una misma especie, en términos de la heterocigosis media por gen (H), vendrá esencialmente determinada por el producto del tamaño eficaz de la población N_e y la tasa de mutación ν por gen y por generación, más que por N_e y ν separadamente. Concretamente, H es igual a $4 N_e\nu / (4 N_e\nu + 1)$. Por ejemplo, si la tasa de mutación es 10^{-6} y el tamaño eficaz de la población es 10^5, la heterocigosis media por gen será aproximadamente 0,286; es decir, el 28,6 por ciento de los individuos, por término medio, serán heterocigotos en cada locus. Cuanto mayor sea el tamaño de la población o la tasa de mutación por gen y por generación, más próximo a la unidad (es decir, al cien por cien) será el valor de la heterocigosis media.

El modelo supone que los alelos se identifican a nivel génico en términos de sustituciones reales de nucleótidos. Sin embargo, la mayoría de las observaciones de la variabilidad dependen de la electroforesis de las proteínas, que tiene mucho menos poder resolutivo y está lejos de revelar todas las sustituciones de nucleótidos (o tan siquiera todos los cambios de aminoácidos); así pues, la heterocigosis observada es menor que la real. Incluso cuando se altera el modelo para tener en cuenta este problema, las poblaciones muy grandes deberían manifestar, según la teoría neutralista, casi un cien por cien de heterocigosis. Cuando las observaciones sugieren algo distinto, la teoría es objeto de críticas. Por ejemplo, Francisco J. Ayala, de la Universidad de California en Davis, ha comunicado que para la mosca neotropical *Drosophila willistoni,* cuyo tamaño poblacional eficaz él estima muy grande, de 10^9, ha observado una heterocigosis de aproximadamente un 18 por ciento. Señala que incluso suponiendo una tasa de mutaciones neutras por generación muy baja, 10^{-7}, la heterocigosis predicha es prácticamente del 100 por cien.

Existen por lo menos dos vías para justificar esta aparente incoherencia. En primer lugar, es posible que el tamaño eficaz de la población de *D. willistoni* no llegara a 10^9 aún cuando el tamaño actual aparente sea enorme. Se puede demostrar matemáticamente que la variabilidad genética debida a los alelos neutros puede verse fuertemente reducida por un "cuello de botella" sufrido, de vez en cuando, por la población, después del cual hacen falta millones de generaciones para que la variabilidad pueda alcanzar de nuevo el nivel teórico característico de una población muy grande mantenida durante un largo período de tiempo. En este sentido, las especies neotropicales, como *D. willistoni,* pueden estar aún bajo el efecto de cuello de botella impuesto por la última glaciación continental, ocurrida entre unos 10.000 y 30.000 años atrás. Además, la eliminación local de colonias de una especie, que puede ser bastante frecuente, puede reducir el tamaño eficaz de la población.

Una segunda posibilidad es que, como Ohta propuso por primera vez en 1973, la mayoría de los alelos "neutros" no sean realmente neutros sino ligeramente deletéreos. Adoptando la idea de Otha, pero dejando también lugar para las mutaciones verdaderamente neutras, he considerado un modelo en el que el coeficiente de selección en contra del mutante, s', sigue una distribución determinada (la distribución gamma) [*véase la ilustración de la página siguiente*]. La tasa de mutación para las formas variantes cuyo valor de selección negativa (s') sea más pequeño que el recíproco del doble del tamaño de la población, o $1/(2N)$, puede considerarse como la tasa de mutación neutra eficaz ν_e. Puede demostrarse que esta tasa de mutación neutra eficaz decrece al aumentar la población; en el caso expuesto dicha tasa es proporcional a 1 dividido por la raíz cuadrada del tamaño de la población. En este modelo, el nivel de heterocigosis crece muy lentamente al aumentar la población. Además, dando un valor supuesto realista al tiempo de generación, la tasa evolutiva en función de las sustituciones mutantes sería aproximadamente constante por año para varias líneas filogenéticas, si fuera constante la tasa de mutación por generación. Téngase en cuenta que aunque esta interpretación hace uso de la selección natural, es completamente distinta de la interpretación seleccionista.

Un enfoque cuantitativo

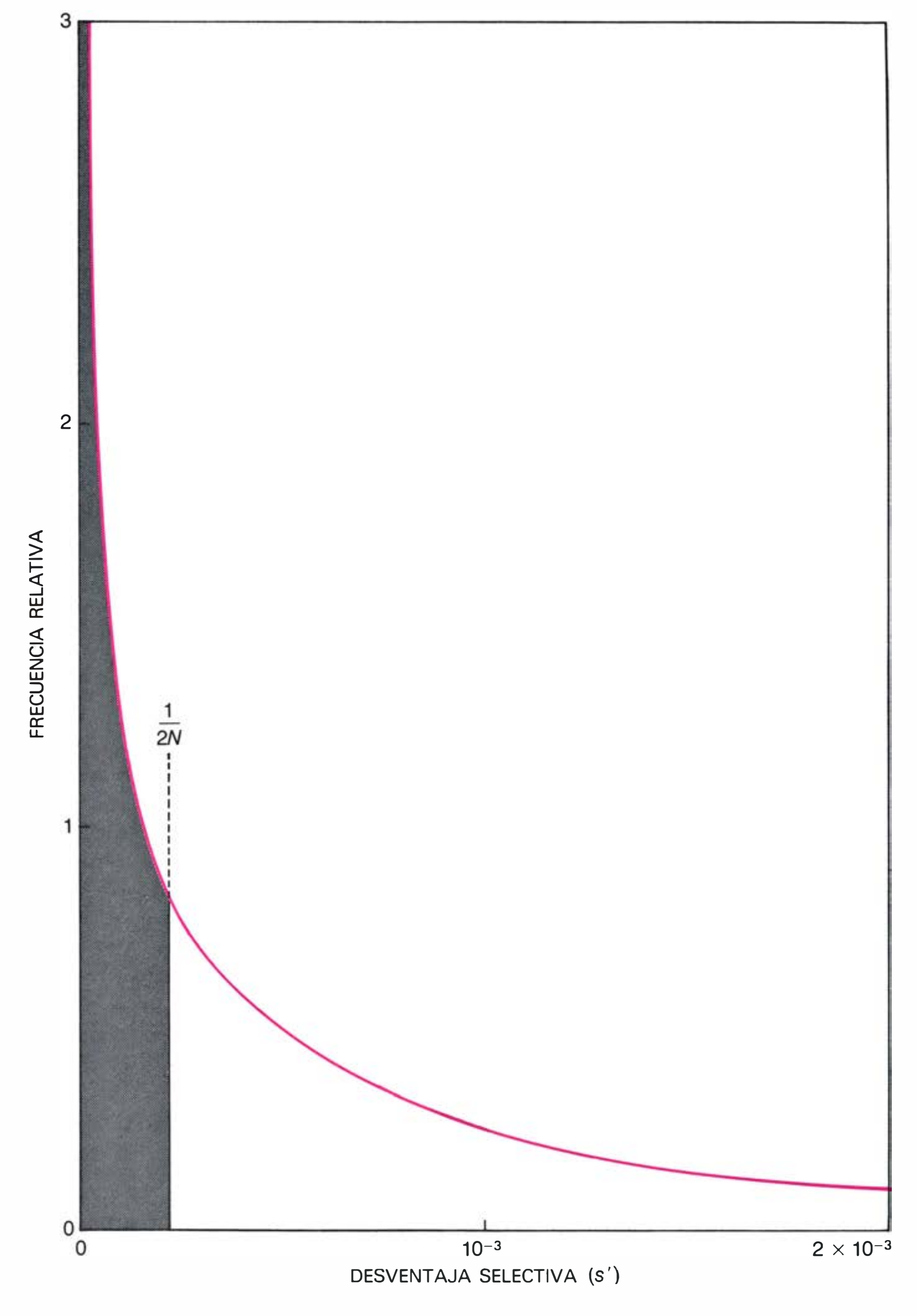
La teoría neutralista de la evolución molecular y del polimorfismo que he de-

sarrollado en colaboración con mis colegas Ohta y Takeo Maruyama se distingue de la mayoría de los enfoques seleccionistas –en particular del enfoque de la genética ecológica– en que apunta a una descripción cuantitativa de la evolución molecular, que intentamos llevar a término utilizando las ecuaciones de difusión. Es una aventura de lo que podría llamarse genética de poblaciones molecular. Nei y sus colaboradores en Houston han contribuido en gran manera a este esfuerzo, en particular relacionando las predicciones teóricas con las observaciones reales. Han demostrado, por ejemplo, que la varianza de la heterocigosis para un determinado enzima de una especie puede predecirse con bastante exactitud por medio de la teoría neutralista sobre la base de las observaciones de la heterocigosis media.

Por ser cuantitativa, nuestra teoría se presta a comprobación y, por tanto, es mucho más susceptible a la crítica, cuando no se cumple, que las teorías seleccionistas, que pueden recurrir a tipos especiales de selección que se ajusten a las circunstancias especiales y que normalmente no sirven para hacer predicciones cuantitativas. Por contra, para comprobar la teoría neutralista, se precisa estimar valores como las tasas de mutación, los coeficientes de selección, los tamaños de las poblaciones y las tasas de migración. Muchos biólogos que estudian la evolución sostienen que tales valores, referentes a la genética de poblaciones, nunca pueden ser determinados con exactitud y que, por consiguiente, cualquier teoría que dependa de ellos no es más que un inútil ejercicio. Pero yo creo que estos valores deben investigarse y medirse si queremos conocer los mecanismos de la evolución. Astrónomos y cosmólogos no pueden prescindir, evidentemente, de las teorías expuestas en función de los valores astronómicos por el mero hecho de que tales valores sean difíciles de calcular con exactitud.

La selección darwiniana actúa principalmente sobre los fenotipos generados por la actividad de muchos genes. Las condiciones ambientales desempeñan, sin duda, un papel decisivo en la determinación de aquellos fenotipos que van a ser seleccionados; a la selección darwiniana, o positiva, poco le importa cómo los fenotipos son determinados por los genotipos. Las leyes que rigen la evolución molecular son claramente distintas de las que rigen la evolución fenotípica. Aun cuando la visión darwinista de la selección natural prevalezca en la determinación de la evolución a nivel fenotípico, a nivel de la estructura interna del material genético gran parte de los cambios evolutivos están impulsados por la deriva genética. Aunque este proceso al azar es lento e insignificante en el marco de la efímera existencia del hombre, en términos de tiempo geológico contribuye al cambio en una gran medida.

Mis colegas me han dicho, directa o indirectamente, que la teoría neutralista no es importante biológicamente porque los genes neutros no están implicados en la adaptación. Mi opinión es que lo importante es encontrar la verdad, y si la teoría neutralista es una hipótesis de investigación válida, entonces establecer la teoría, comprobarla y defenderla es una tarea científica que merece la pena.



EL MODELO TEORICO supone que la mayoría de los alelos "neutros" son en realidad ligeramente deletéreos. Por tanto, se hallan sometidos a un pequeño coeficiente de selección negativa *s'*. El modelo pone de manifiesto la distribución de frecuencias (*curva de color*) del valor *s'* para las mutaciones en diversos puntos de un gen. Las mutaciones cuya desventaja selectiva es menor que la unidad dividida por el doble del tamaño de la población *N* son neutras en realidad. El área englobada bajo la curva (*gris*) que ocupan dichos mutantes decrece al aumentar la población. Al disminuir, así, la fracción de mutantes realmente neutros y al estar expuestos en consecuencia más mutantes a la selección negativa, la tasa de evolución molecular se ve reducida. En este modelo, el nivel de heterocigosis (una medida de la variabilidad) aumenta de una forma lenta con el tamaño de la población, poniendo en justo acuerdo las predicciones de la teoría neutralista (avanzada y sostenida por el autor) con las observaciones.

Las galaxias primitivas

Las primeras galaxias que se formaron tras la "gran explosión" no han sido vistas, pero hay razones para creer que puedan serlo. Las características de galaxias más antiguas indican cuáles pudieron ser las de aquellas más jóvenes

David L. Meier y Rashid A. Sunyaev

Por ser finita la velocidad de la luz, el astrónomo puede en principio mirar hacia atrás en el tiempo casi hasta la misma creación del universo con sólo observar objetos tan distantes que su luz haya tardado 16.500 millones de años, la mejor estima actual de la edad del universo, en alcanzar el sistema solar. En particular, existe la intrigante posibilidad de que a esas vastas distancias lleguemos a contemplar galaxias en proceso de formación. Como se cree que la mayoría de las galaxias se han formado de la misma manera, la investigación de tales galaxias primitivas podría ayudar a explicar de qué manera se constituyó nuestra propia galaxia hace 13.000 millones de años. La búsqueda de galaxias primitivas no se ha visto todavía coronada por el éxito, principalmente porque debe ser difícil distinguirlas de otros objetos débilmente luminosos, como los quasars. Además, habrán de hallarse tan lejos, que muchas resultarán, quizá, demasiado débiles para apreciarlas con los telescopios ordinarios.

La situación debe mejorar hacia 1985, época en la que está previsto el lanzamiento del superpotente telescopio espacial mediante la Lanzadera Espacial. Dotado de una resolución 10 veces mayor que la del mejor instrumento óptico ahora existente, el telescopio espacial estará capacitado para detectar objetos estelares 100 veces más débiles que cualquiera de los que se han detectado hasta ahora. Aun los objetos difusos, tales como las galaxias distantes, deben resultar visibles a distancias mucho mayores que las actualmente alcanzadas. Nuestro propio trabajo se ha centrado en determinar las propiedades de una galaxia primitiva, por ejemplo su espectro y la naturaleza de su imagen fotográfica, de manera que los astrónomos puedan identificar una cuando la vean.

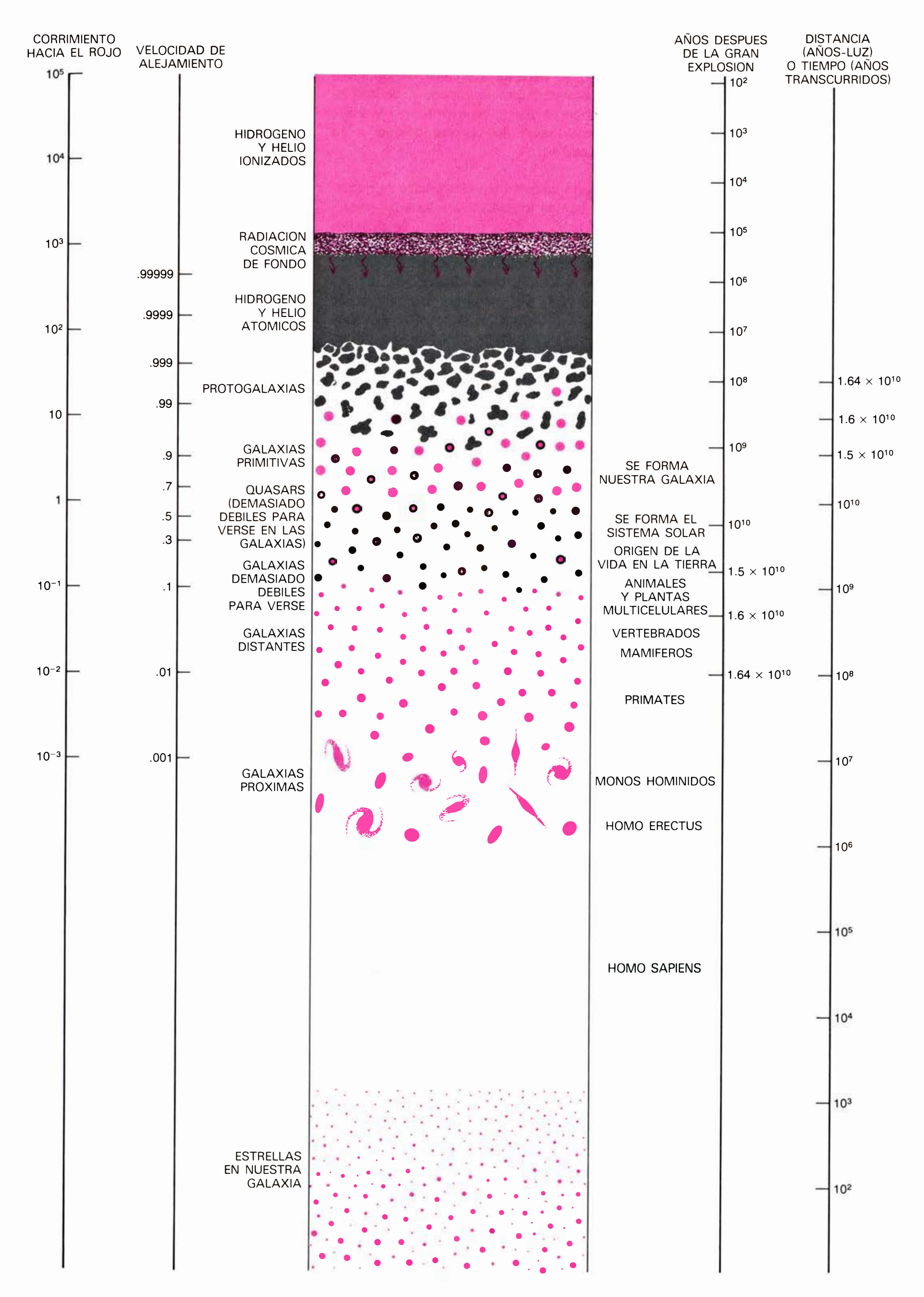
De acuerdo con la teoría cosmológica más aceptada, el universo comenzó con una explosión a partir de un estado superdenso, en el cual la velocidad de expansión aumenta con la distancia al observador. La longitud de onda a la cual la radiación electromagnética procedente de un objeto distante alcanza a la Tierra aumenta a causa de la velocidad de separación del objeto respecto al observador. Este es el bien conocido corrimiento hacia el rojo, llamado así porque si la radiación se encuentra en la región visible del espectro, el efecto tiende a hacerla más rojiza. La cuantía del corrimiento hacia el rojo es una medida no solamente de la lejanía del objeto, sino también, puesto que estamos mirando hacia atrás en el tiempo, de su edad desde la "gran explosión".

Cuanto menor sea el corrimiento hacia el rojo, tanto más joven es la galaxia. Las galaxias más próximas, comparables en edad a la nuestra, se encuentran a distancias de unos 20 millones de años-luz y se están alejando con un corrimiento hacia el rojo de 0,001, o bien 0,1 por ciento. Por encima de un corrimiento hacia el rojo de 0,1 por cien, entran en escena los quasars. Esos antiguos objetos semejantes a estrellas son corrientes para corrimientos hacia el rojo próximos a 2,5. Se cree que los quasars existen en el centro de galaxias, que, por lo demás, son normales. Esto sugiere que las galaxias se debieron formar antes que los quasars, de modo que las galaxias primitivas habrán de manifestar un corrimiento hacia el rojo superior a 2,5. Se puede determinar un límite superior del corrimiento hacia el rojo de una galaxia en el proceso de formación extrapolando hacia atrás desde el universo actual hasta la época en que las galaxias se tocaban unas con otras. Esto ocurre para un corrimiento hacia el rojo de aproximadamente 100, que representa la época más temprana posible en que las galaxias pudieron existir como entidades separadas.

La búsqueda de galaxias primitivas para corrimientos hacia el rojo de entre 2,5 y 100 es una empresa extremadamente difícil. De hecho, nadie ha propuesto todavía un solo candidato convincente para galaxia primitiva. La razón es que las galaxias primitivas deben ser sumamente débiles a tan grandes distancias. La galaxia más distante detectada hasta ahora se encuentra a un corrimiento hacia el rojo de sólo 0,75. Es incluso posible que el telescopio espacial no sea capaz de detectar galaxias normales a los altos corrimientos hacia el rojo característicos de las galaxias primitivas. Hay muchas indicaciones, sin embargo, de que las galaxias poseen muchas más estrellas brillantes en el proceso de formación que en cualquier otra fase de su ciclo vital. Ello significa que serían mucho más brillantes que las galaxias normales para el mismo corrimiento hacia el rojo. La consecuencia es notable: algunas de las galaxias primitivas y quizá todas ellas podrían estar dentro del alcance de los instrumentos actuales.

La escasa luminosidad de las galaxias primitivas las hace, de todas maneras, difíciles de distinguir de otros objetos débiles tales como los quasars. Muchos fenómenos astronómicos, cualquiera que sea su naturaleza y su distancia, tienen un aspecto muy parecido en las placas fotográficas. Las fotografías que registran objetos débiles recogen estrellas dis-

HISTORIA DEL UNIVERSO, presentada aquí en escala logarítmica. Puede verse a través de los telescopios porque la velocidad de la luz es finita. En principio, el astrónomo puede mirar hacia atrás en el tiempo casi hasta la creación del universo por el procedimiento de observar objetos tan distantes que su luz ha tardado 16.500 millones de años, es decir, la edad del universo, en llegar a nuestra galaxia. La lejanía de un objeto se mide por su corrimiento hacia el rojo: la velocidad de alejamiento del objeto dividida por la velocidad de la luz. Aquí la velocidad de alejamiento se expresa como fracción de 1 a velocidad de la luz. En la parte izquierda del cuadro central se relacionan los objetos astronómicos que predominan para varios corrimientos hacia el rojo. A la derecha de la figura, y al objeto de ilustrar el efecto de la "mirada hacia atrás", se relacionan algunos acontecimientos en la historia de la Tierra y su evolución biológica que tuvieron lugar en el instante correspondiente de la "mirada hacia atrás".



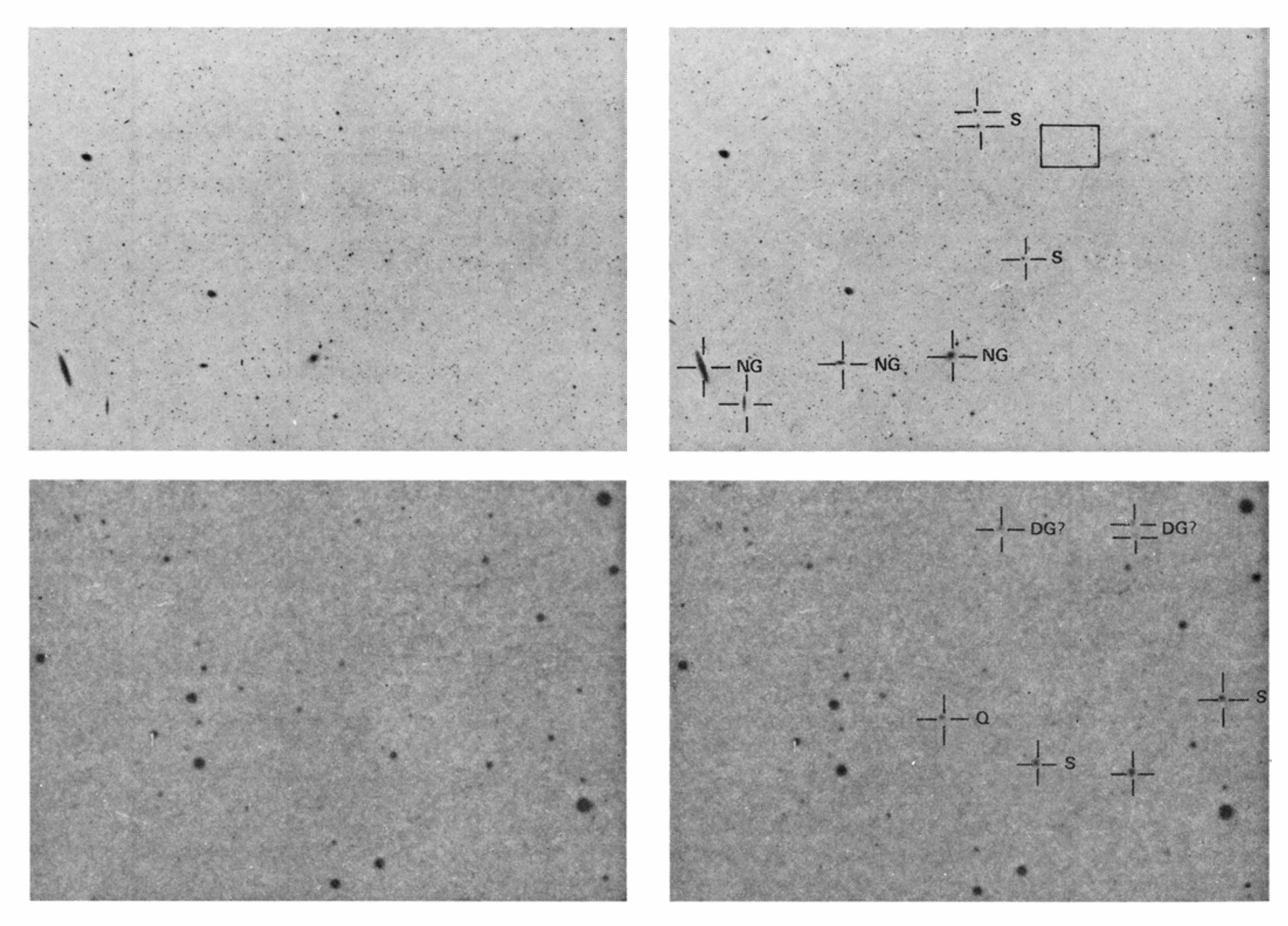
CORRIMIENTO HACIA EL ROJO
10^5
10^4
10^3
10^2
10
1
10^{-1}
10^{-2}
10^{-3}
VELOCIDAD DE ALEJAMIENTO
.99999
.9999
.999
.99
.9
.7
.5
.3
.1
.01
.001
HIDROGENO Y HELIO IONIZADOS
RADIACION COSMICA DE FONDO
HIDROGENO Y HELIO ATOMICOS
PROTOGALAXIAS
GALAXIAS PRIMITIVAS
QUASARS (DEMASIADO DEBILES PARA VERSE EN LAS GALAXIAS)
GALAXIAS DEMASIADO DEBILES PARA VERSE
GALAXIAS DISTANTES
GALAXIAS PROXIMAS
ESTRELLAS EN NUESTRA GALAXIA
SE FORMA NUESTRA GALAXIA
SE FORMA EL SISTEMA SOLAR
ORIGEN DE LA VIDA EN LA TIERRA
ANIMALES Y PLANTAS MULTICELULARES
VERTEBRADOS
MAMIFEROS
PRIMATES
MONOS HOMINIDOS
HOMO ERECTUS
HOMO SAPIENS
AÑOS DESPUES DE LA GRAN EXPLOSION
10^2
10^3
10^4
10^5
10^6
10^7
10^8
10^9
10^{10}
1.5×10^{10}
1.6×10^{10}
1.64×10^{10}
DISTANCIA (AÑOS-LUZ) O TIEMPO (AÑOS TRANSCURRIDOS)
1.64×10^{10}
1.6×10^{10}
1.5×10^{10}
10^{10}
10^9
10^8
10^7
10^6
10^5
10^4
10^3
10^2

tantes en nuestra galaxia, quasars distantes y galaxias normales asimismo alejadas. En las placas fotográficas sensibles, frecuentemente sólo se pueden distinguir las galaxias de las estrellas y quasars por sus imágenes mayores y algo más difusas, y cuanto menos luminoso sea el objeto, más difícil resultará la distinción. La mejor manera de identificar los objetos suele ser por medio de las características generales de su espectro. Los quasars tienen líneas de emisión, no así las galaxias normales; las galaxias tienen corrimientos hacia el rojo, de lo que carecen las estrellas.

Ninguna de esas distinciones, sin embargo, puede servir para identificar una galaxia en el proceso de formación. Las estrellas masivas, brillantes y calientes que es de esperar existan en las galaxias primitivas ionizarían el gas existente en el medio interestelar, de modo que el espectro de una tal galaxia incluiría líneas de emisión. Es concebible, entonces, que se tomaran las galaxias primitivas por quasars a causa de sus intensas líneas de emisión y grandes corrimientos hacia el rojo. Se necesita más información sobre el espectro que se supone debe corresponder a una galaxia primitiva y sobre cómo difiere del correspondiente a un quasar.

Para determinar el espectro supuesto hemos tomado como base los modelos teóricos de la formación de galaxias. Como muchos factores en tales modelos son desconocidos, los modelos constituyen sólo una tosca descripción del objeto real. Sin embargo, suministran la única información de la que el astrónomo puede fiarse hasta el feliz día en que se identifique definitivamente una galaxia primitiva. Tales modelos se basan en resultados que provienen de muchas áreas diferentes de la astrofísica y la cosmología. Incorporan los procesos físicos de los que se sabe que rigen una galaxia: los macroscópicos de dinámica estelar y dinámica de gases, que determinan el tamaño y forma generales de la galaxia, y los microscópicos de la formación estelar, evolución y muerte que determinan las propiedades de las estrellas constituyentes de la galaxia. Aparte de la semejanza observacional entre las galaxias primitivas y los quasars, las dos clases de objetos pueden estar relacionadas en un sentido más profundo. Los quasars son físicamente diferentes de las galaxias, pero como hemos mencionado, parecen existir en galaxias y eran más abundantes en la infancia del universo. Como resultado, probablemente estén relacionados de alguna manera con la formación de las galaxias.

La teoría más ampliamente aceptada de formación de las galaxias es la teoría de la inestabilidad gravitatoria, que sostiene que las galaxias se condensaron a partir del fluido cosmológico caliente que se expandió a raíz de la gran explosión. Si una región del primitivo universo hubiera tenido casualmente una densidad más alta que la de las regiones



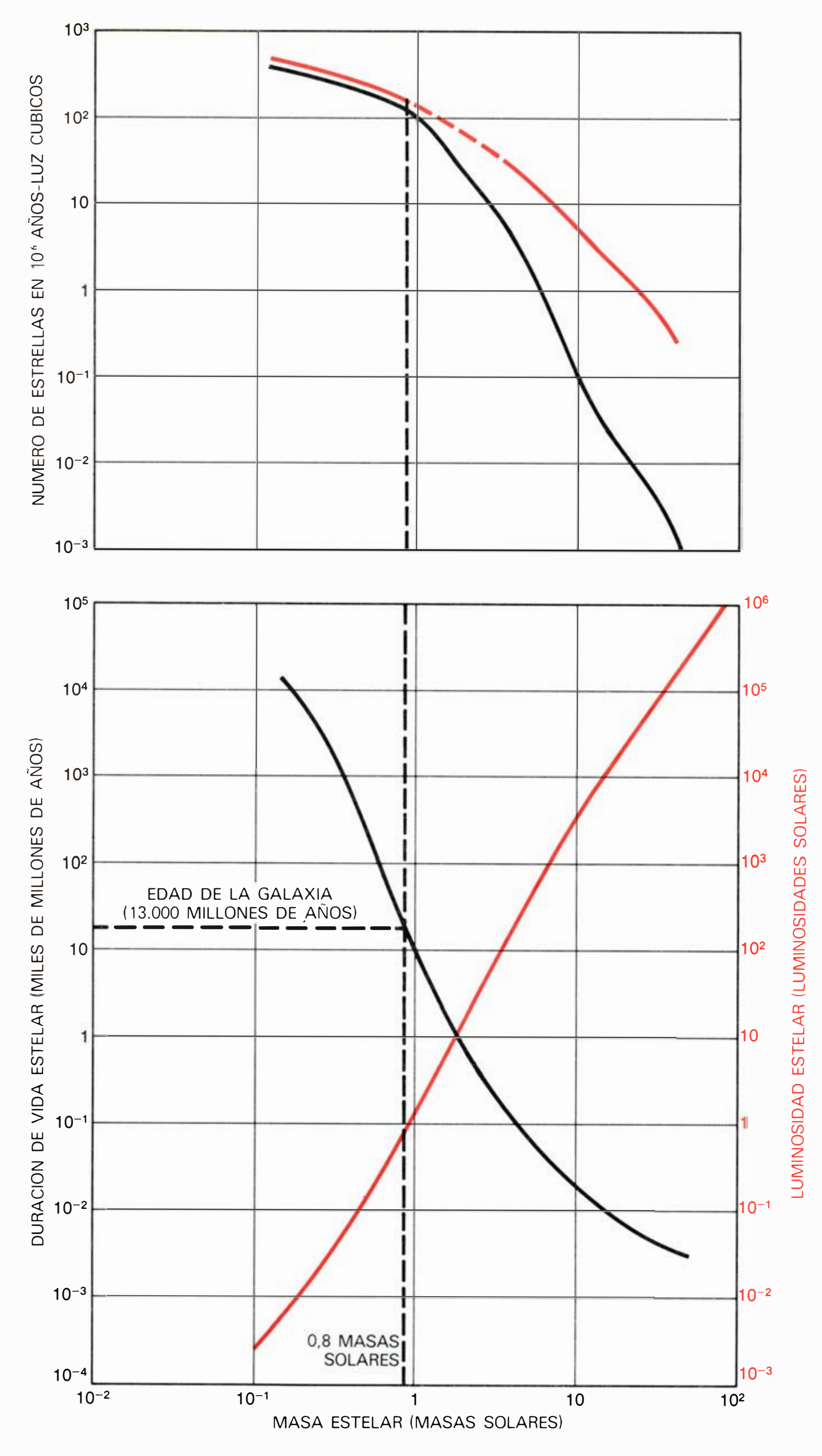
DEBIL LUMINOSIDAD de las primitivas galaxias, que hace difícil distinguirlas de otros objetos débiles tales como los quasars. La fotografía negativa de la parte superior izquierda, hecha con el telescopio Schmidt de 1,2 metros del Monte Palomar, corresponde a una región relativamente grande del cielo (1,5 por 2 grados) en el borde extremo del cúmulo de galaxias de Virgo, a 45 millones de años-luz de nuestra galaxia. En el duplicado de la fotografía en la parte superior derecha se indica la posición de algunas galaxias próximas (*NG*) y de algunas estrellas bastante brillantes (*S*) de nuestra galaxia. La fotografía registra también una multitud de objetos extremadamente tenues. Posiblemente algunos de ellos son galaxias primitivas. La mayoría de las galaxias primitivas, sin embargo, son probablemente demasiado débiles para poder ser vistas. Las fotografía en la parte inferior izquierda corresponde a la parte enmarcada (ocho por 11 minutos de arco) de la fotografía de arriba. En el duplicado fotográfico de la parte inferior derecha se marcan las posiciones de varias estrellas débiles, simbolizadas por *S*, un quasar (*Q*) y candidatos para galaxias distantes (*DG?*).

circundantes, sería atraída gravitatoriamente más hacia sí misma que hacia la materia del ambiente. Si tal región, caracterizada como una perturbación, estuviera libre de presión del gas o de radiación, se contraería bajo su propia gravedad, y por tanto aumentaría su densidad. Este proceso de contracción crearía "gotículas" en un universo que antes era por completo homogéneo.

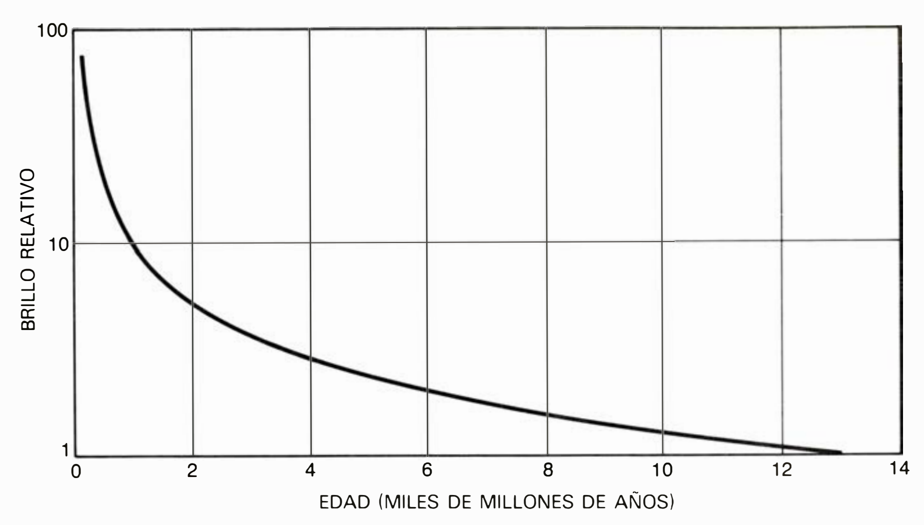
La escala de las perturbaciones variaba, probablemente, desde la masa de un cúmulo globular de estrellas (10^6 veces la masa del Sol) a la masa de un gran agregado de galaxias (10^{15} masas solares). Las perturbaciones que son del tamaño de las galaxias (10^{11} masas solares) se llaman protogalaxias. Inicialmente, cada protogalaxia se expandiría con el resto del universo, pero lo haría a un ritmo ligeramente menor. Después de unos pocos cientos de millones de años, cesaría de expandirse aun cuando el universo continuase haciéndolo. En ese punto, la protogalaxia quedaría en efecto desprendida del universo, libre para colapsar sobre sí misma y convertirse en una galaxia.

¿Qué queremos indicar cuando decimos que se ha formado una galaxia? Para que una protogalaxia se convierta en galaxia deben ocurrir dos cosas. Primero, debe formarse una población de estrellas a partir del gas protogaláctico. Segundo, el gas y las estrellas deben reunirse para formar la bien ordenada estructura de una galaxia. El segundo proceso se conoce con cierta perfección. Con un ordenador de alta velocidad es relativamente sencillo seguir el colapso de un modelo de protogalaxia resolviendo las ecuaciones del movimiento aplicadas a las estrellas y el gas bajo la influencia de su mutua atracción gravitatoria. No se sabe mucho, sin embargo, acerca de cómo se formaron en realidad las estrellas a partir del gas en el curso de su colapso. Y lo poco que se sabe proviene de observaciones de nuestra galaxia y las galaxias próximas. Tales observaciones sugieren posibles acontecimientos que podrían desencadenar la formación de estrellas y también la proporción de estrellas de diferentes masas que se crean en un brote de formación de estrellas. Con estas observaciones es posible desarrollar una descripción semicuantitativa, generada por ordenador, de la formación de las galaxias.

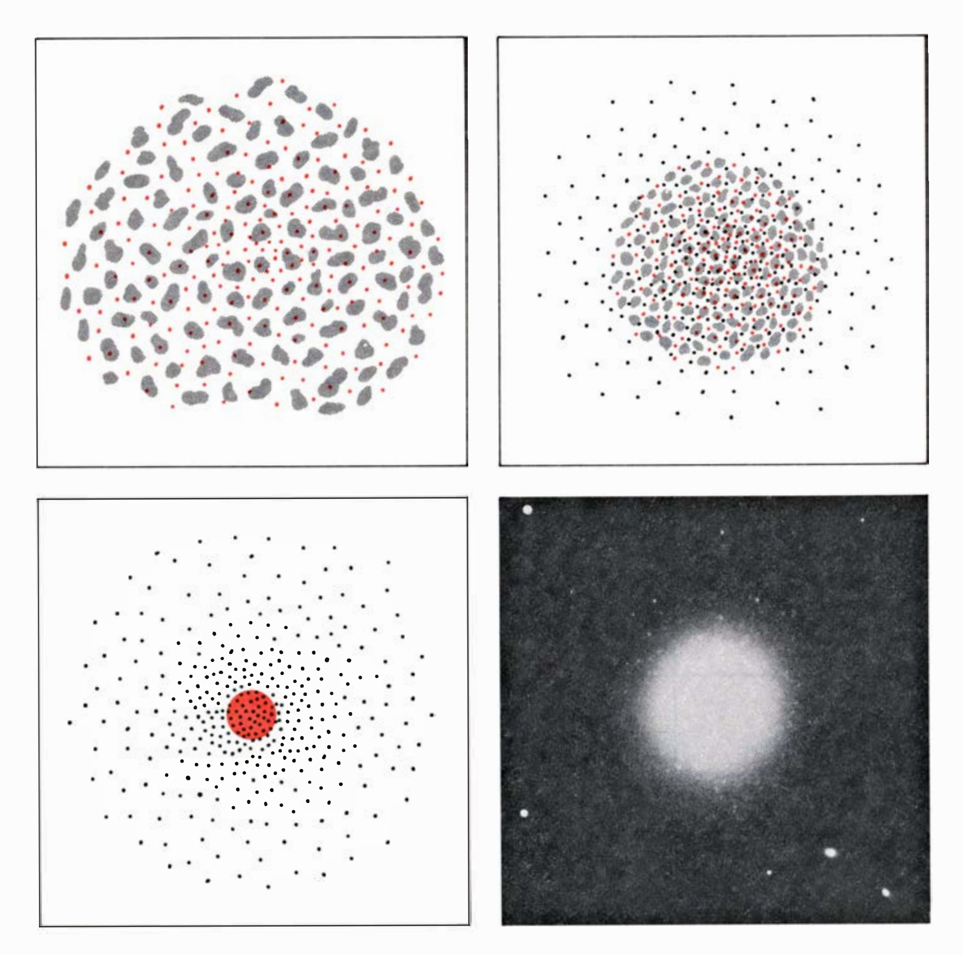
Como las galaxias, las estrellas probablemente se forman del gas interestelar a causa de inestabilidades gravitatorias. A diferencia de la formación de galaxias, sin embargo, el nacimiento de una estrella no suele ser un proceso espontáneo, porque la presión del gas inte-



FUNCION INICIAL DE MASA, o distribución de estrellas de diferentes masas que se formaría en un brote de formación de estrellas en una galaxia. Dicha función aparece representada en el diagrama de la parte superior. La curva negra muestra la distribución actual de estrellas según sus masas en la vecindad del Sol. Las estrellas con masas de 0,8 veces la masa del Sol tienen duraciones de vida mayores que la edad de la galaxia, de modo que la distribución de masas de esas estrellas es la misma que era inicialmente. Cuanto mayor la masa de la estrella, sin embargo, tanto más rápidamente se consume (*curva negra en la parte inferior*). Las estrellas con masas superiores a 0,8 veces la del Sol tienen duraciones de vida menores que la edad de la galaxia, de manera que las que se formaron hace 13.000 millones de años, cuando se creó la galaxia, se han extinguido. Para masas por encima de 0,8 masas solares, la función inicial de masa (*curva en color de arriba*) se puede determinar extrapolando hacia atrás en el tiempo, reemplazando las estrellas extinguidas por otras vivientes. Aunque el número de estrellas decrece al aumentar la masa, el decrecimiento no es suficiente para compensar el aumento de luminosidad estelar con la masa (*curva coloreada de abajo*). La luminosidad de un cúmulo de estrellas recién formado es bastante alta y se debe principalmente a las brillantes estrellas masivas de vida corta que integran el cúmulo en cuestión. Una galaxia primitiva debe ser, pues, muy brillante.



LUMINOSIDAD DE UN CUMULO DE ESTRELLAS. Dicha luminosidad es máxima cuando se forma a causa de la multitud de brillantes estrellas de gran masa. A medida que estas estrellas de vida corta se van extinguiendo, la luminosidad del cúmulo estelar disminuye rápidamente. El "brillo relativo" se determina dividiendo la luminosidad de una población de estrellas por la luminosidad a la edad de 13.000 millones de años, que es la edad de nuestra galaxia. El brillo relativo de una población de estrellas constituye una medida de cuántas veces brillaba más en su origen que en la actualidad.



FORMACION DE UNA GALAXIA, según algunos de los modelos realizados con ordenador. Comienza con un cúmulo de pequeñas nubes de gas de gran masa, de una extensión de unos 200.000 años-luz (*parte superior izquierda*), que chocan, colapsan y forman una población de estrellas brillantes (*puntos coloreados*). A medida que las nubes caen hacia dentro comprimiéndose en un volumen menor, a raíz de su mutua atracción gravitatoria, aumenta la velocidad de formación de estrellas. Después de unos 200 millones de años, la protogalaxia tiene un diámetro de unos 100.000 años-luz (*parte superior derecha*) y la primera población de estrellas (*puntos negros*) se ha extinguido. Las nubes de gas continúan contrayéndose a ritmo creciente. Después de unos 300 millones de años, el gas ha quedado confinado en el centro de la galaxia, conduciendo a un máximo en la velocidad de formación de estrellas (*parte inferior izquierda*). El denso núcleo de estrellas de la galaxia tiene un diámetro quizá no superior a 10.000 años-luz. En esta fase, la galaxia es uno de los objetos más brillantes del universo. Desde una gran distancia, sin embargo, hasta un objeto tan luminoso aparecería como una estrella débil. Después de esta fase, la formación de estrellas cesa porque se ha agotado el gas de la galaxia. A medida que las estrellas más brillantes se extinguen, la galaxia luce con menos brillo pero en forma más constante, apareciendo como una gigantesca estructura esférica, como la galaxia M87 (*parte inferior derecha*).

restelar es usualmente suficiente para impedir que perturbaciones tan pequeñas como las estrellas colapsen. Si se comprime el gas por algún fenómeno violento, la densidad puede aumentar hasta el punto en que la gravedad venza la presión del gas; la perturbación se contrae para formar una estrella. En nuestra galaxia, los fenómenos violentos incluyen una ola espiral de densidad de la clase responsable de la estructura a modo de rueda de fuegos artificiales de las galaxias de disco y también ondas de choque producidas por las explosiones de supernovas y por las regiones de gas ionizado en expansión que rodean a las estrellas calientes de gran masa. En otras galaxias, la formación de estrellas puede también haber sido desencadenada por choques o casi choques con galaxias vecinas. En una protogalaxia, el violento movimiento de hundimiento hacia dentro es en sí mismo probablemente el principal proceso responsable de la formación de estrellas. Una protogalaxia se puede considerar como un sistema de nubes gaseosas moviéndose en órbitas a gran velocidad y dando nacimiento a estrellas cuando las nubes chocan.

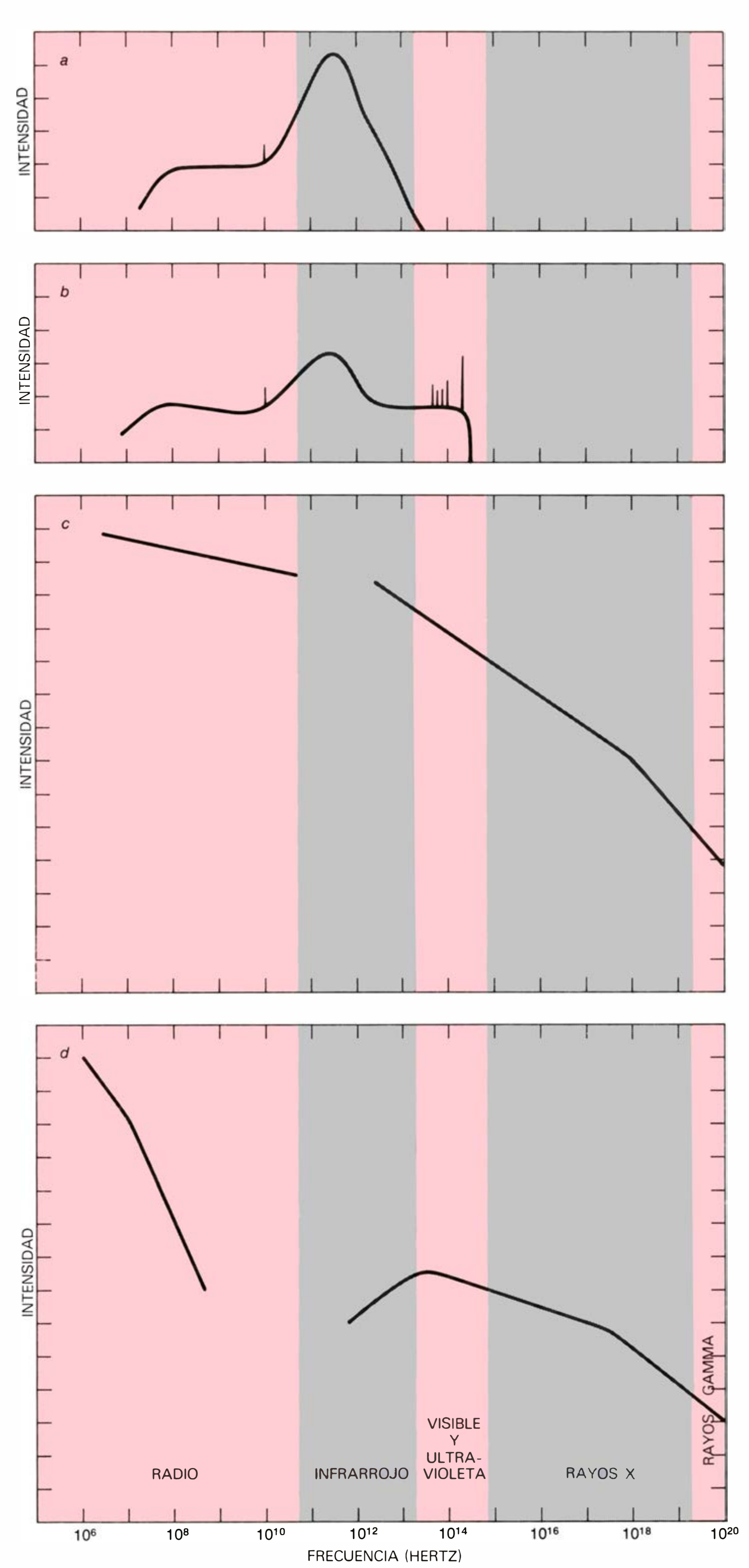
Las observaciones de estrellas en la vecindad del Sol han hecho posible estimar la distribución de estrellas de diferentes masas que se crearían en un brote de formación de estrellas. Tal distribución se llama función inicial de masas. La función local de masas en el momento presente –la distribución de masas de las estrellas en la vecindad del Sol– comprende muchas estrellas que pesan menos de 0,8 veces la masa solar, pero sólo unas pocas con masa superior a aquélla. Las estrellas más masivas son escasas porque tales estrellas se consumen más deprisa y de ahí que vivan menos tiempo. El Sol vivirá durante 10.000 millones de años, pero una estrella que tenga una masa de sólo diez veces la del Sol brillará con una luminosidad 20.000 veces mayor y vivirá durante 10 millones de años, es decir, una milésima parte de la vida del Sol. Las estrellas de masa superior a 0,8 masas solares tienen tiempos de vida que son más cortos que la edad de la galaxia, de modo que las estrellas que se formaron cuando se creó la galaxia hace 13.000 millones de años se han extinguido ya.

Las estrellas con masas inferiores a 0,8 masas solares gozan de tiempos de vida más largos que la edad de la galaxia. Las estrellas que se encuentran en este dominio de masas menores han ido aumentando en número en el curso de los muchos brotes de formación de estrellas, de modo que la distribución rela-

tiva de masas de las estrellas debe ser idéntica a la de las estrellas creadas en cualquier otro brote, admitiendo naturalmente que las estrellas se han formado siempre de la misma manera. En otras palabras, por debajo de 0,8 masas solares, la función local de masas en la actualidad es la función de masas inicial. Por encima de 0,8 masas solares, la función de masas inicial se puede determinar extrapolando hacia atrás en el tiempo mediante la sustitución de estrellas muertas por otras vivientes. La relación resultante, que fue obtenida por Edwin E. Salpeter de la Universidad de Cornell, indica que cuando se crea un grupo de estrellas, por cada vez que se multiplique por 10 la masa se divide por 22 el número de estrellas.

Si la relación de Salpeter vale para todos los episodios de formación de estrellas, resulta tener sorprendentes implicaciones en relación con la formación de galaxias y con la búsqueda de las galaxias primitivas. Aunque el número de estrellas decrece al aumentar la masa, el decrecimiento no es suficiente para compensar el aumento de la luminosidad estelar con la masa (un factor aproximado de 10.000 por cada 10 veces que aumente la masa). Esto significa que la luminosidad del cúmulo de estrellas recién formado es muy alta y se debe, sobre todo, a las estrellas de gran masa, muy brillantes y de vida corta. Transcurrido poco tiempo, sin embargo, las estrellas brillantes se extinguen y la luminosidad del cúmulo decrece rápidamente. Como resultado, una galaxia primitiva, en la que es de esperar que el ritmo de formación de estrellas sea mucho más rápido que el de las actuales galaxias, debe ser considerablemente brillante.

Richard B. Larson, de la Universidad de Yale, ha desarrollado complicados modelos para ordenador de tres tipos de galaxia —esféricas, elípticas y de disco— en el proceso de colapso. La galaxia esférica es la más brillante. Se trata, por tanto, del candidato más firme para galaxia primitiva detectable. Después que la galaxia fotosférica alcanza su máxima



LOS ESPECTROS de las estrellas de gran masa en varias fases de su ciclo de vida contribuirán al espectro de una galaxia primitiva. Las estrellas masivas envueltas por nubes de polvo emiten radioondas que atraviesan el polvo (*a*). El polvo absorbe radiación de las longitudes de onda mayores, pero reemite en forma de emisión térmica en el infrarrojo. El monóxido de carbono en las nubes de moléculas contribuye con una marcada línea de emisión a 115 gigahertz (miles de millones de ciclos por segundo). La intensidad de la radiación visible y ultravioleta es mucho mayor para estrellas en que el polvo envolvente se ha dispersado o destruido (*b*). Una estrella masiva que explota deja un resto que emite radioondas y rayos X (*c*). Los pulsars emiten radioondas (*d*).

expansión, su gas comienza a contraerse, y sus pequeñas nubes internas se aceleran, chocan y forman estrellas. La mayoría de esas estrellas continuarán luciendo durante miles de millones de años y sus órbitas serán responsables de la forma general de la galaxia. Las estrellas masivas, sin embargo, pronto explotan violentamente, lanzando al gas interestelar los núcleos de elementos pesados sintetizados por reacciones nucleares en su interior.

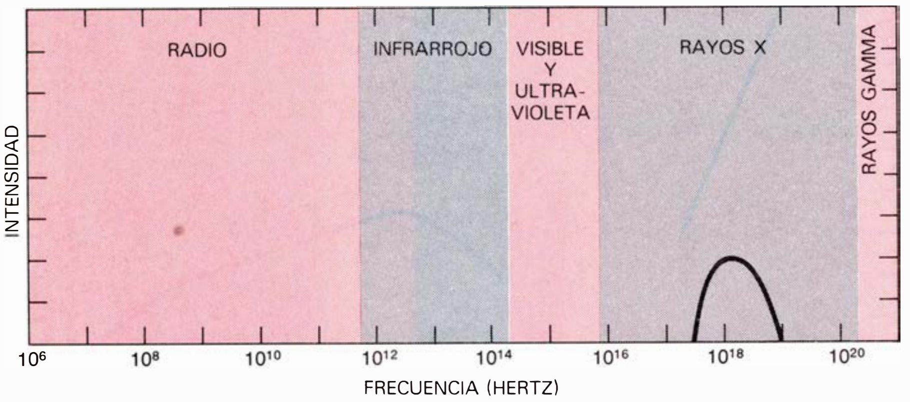
Una vez se han formado las estrellas, su comportamiento dinámico difiere del comportamiento del gas. Las órbitas de las estrellas simplemente se adaptan a la forma en contracción de la galaxia. Las nubes de gas, sin embargo, tienden a encontrarse con otras nubes de gas, de modo que sus órbitas degeneran y las nubes se precipitan hacia el centro de la galaxia. Esto da como resultado que el ciclo completo de encuentro de nubes, formación de estrellas, explosión estelar y enriquecimiento en elementos pesados continúe de un modo acelerado. Con el tiempo, el gas queda confinado en el centro de la galaxia, lo que conduce a un máximo en la velocidad de formación de estrellas. A distancia de unos pocos miles de años-luz del centro, miríadas de estrellas lucen brillantemente y explotan, el gas se hace luminoso y el polvo de elementos pesados radia copiosamente en el infrarrojo. En esta fase, la galaxia es uno de los más brillantes objetos en el universo, cientos de veces más luminoso que las galaxias actuales.

Después, la actividad de la galaxia cesa bruscamente. El gas se ha agotado; ya no queda más para colapsar formando estrellas. Las restantes estrellas brillantes se agotan muy pronto. A medida que las estrellas con duración de vida promedio comienzan a morir, dejando solas a las más antiguas y menos brillantes, la galaxia luce de manera menos brillante pero más constante en forma de gigantesca estructura esférica.

Tales modelos de formación de galaxias sugieren que las galaxias primitivas nacieron en su mayor parte entre 100 millones y algunos miles de millones de años después de la gran explosión. Esto las coloca en corrimientos hacia el rojo entre aproximadamente 3 y 30, lo que concuerda con la estima que anteriormente hicimos a base de los corrimientos hacia el rojo de los quasars. Con los telescopios actuales, la mayoría de las galaxias primitivas con corrimientos hacia el rojo menores apenas si serían visibles, y las que los tuvieran mayores resultarían completamente invisibles. Las galaxias primitivas excepcionalmente brillantes, sin embargo, deben aparecer con claridad. El telescopio espacial quizá logre detectar todas las galaxias primitivas, si exceptuamos las de máximo corrimiento hacia el rojo.

El modelo también sugiere que la relación del brillo del núcleo de la galaxia al brillo de sus regiones periféricas es mayor para una galaxia primitiva que para una galaxia normal. Esto significa que la imagen de una galaxia primitiva se asemeja a la imagen de las estrellas y los quasars; por ello es comprensible que sean difícilmente discernibles. A una distancia de 16.000 millones de años-luz, el brillante núcleo, aunque tendría un diámetro de miles de años-luz, mostraría un diámetro aparente de sólo un segundo de arco. El telescopio espacial podrá resolver con claridad la estructura de tal objeto. De hecho, en un grado cuadrado de cielo deberán poderse detectar, con el tiempo, miles de galaxias primitivas.

Trabajando con Beatrice M. Tinsley, de la Universidad de Yale, emprendimos el cálculo del espectro probable de una galaxia primitiva. Logramos obtener mucha información de nuestra propia galaxia, como había sido el caso cuando la determinación de la función inicial de masa de las estrellas en el universo. En una galaxia primitiva, las estrellas recién nacidas y los objetos asociados con ellas deben contribuir al espectro de la galaxia. Aunque esas estrellas y objetos son distantes y antiguos, es de esperar que resulten casi idénticos a

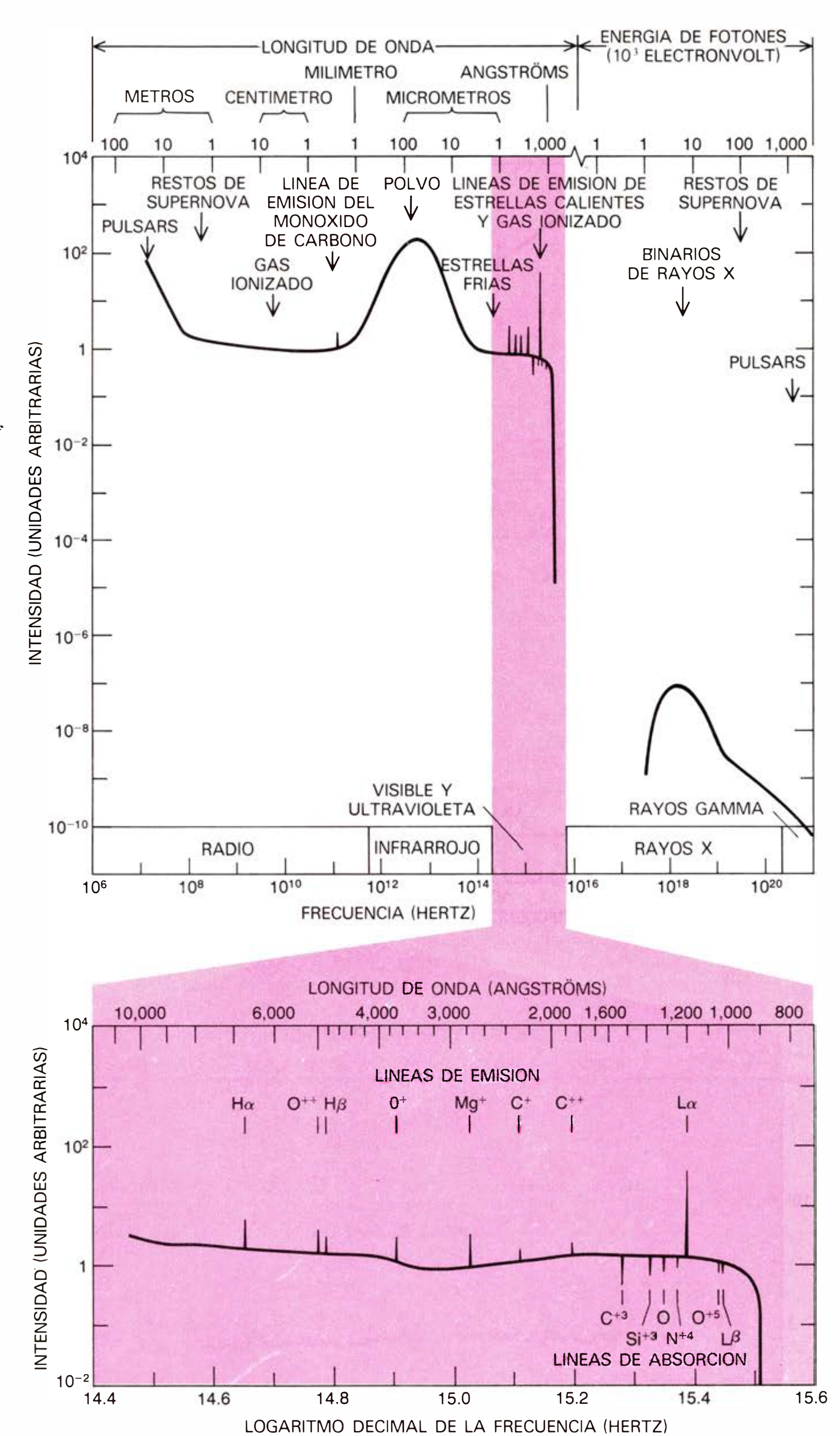


ESTRELLAS DE NEUTRONES Y AGUJEROS NEGROS. Unas y otros constituyen los compactos residuos de una explosión de supernova. Podrían contribuir al espectro de una galaxia primitiva en una forma poco corriente. Si la estrella que explotó era originariamente un miembro de un sistema binario en el que la otra estrella tenía una duración de vida mayor, la estrella de neutrones o el agujero negro podría quedar en órbita alrededor de la estrella de vida más larga (*arriba*). Se podría entonces transferir materia a este compacto objeto mediante un "viento" estelar o por gravedad, con lo que el objeto arrancaría materia de la estrella. Este proceso daría lugar a una fuerte emisión de rayos X (*abajo*).

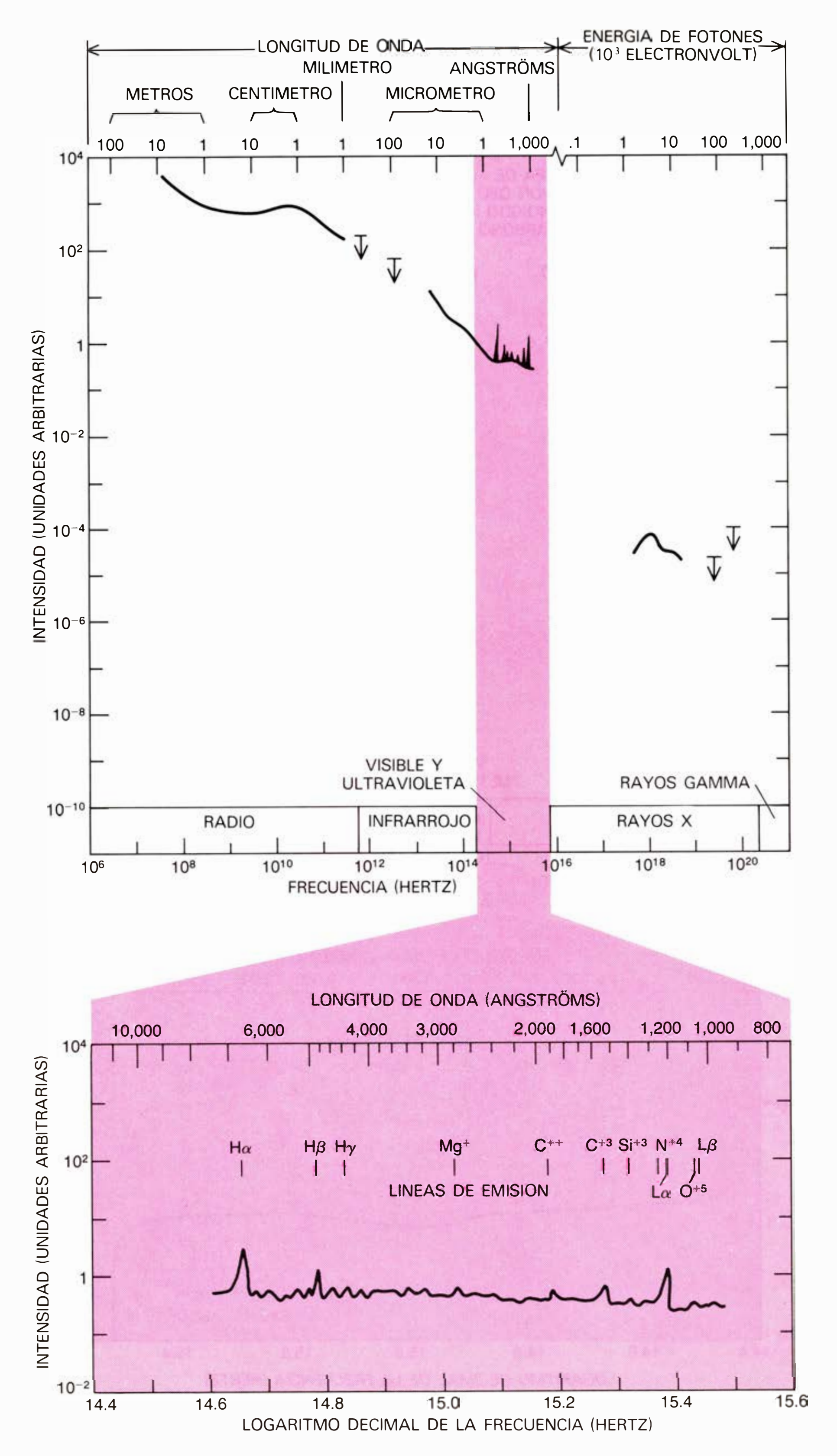
las estrellas y objetos de nuestra galaxia. Esta estrecha semejanza es muy verosímil, porque el modelo indica que los elementos pesados se crean y distribuyen durante la fase inicial del colapso de una protogalaxia, dando a la estrella recién nacida la misma composición química, y por tanto el mismo espectro, que una estrella recién nacida en nuestra galaxia. Como resultado, todos los ingredientes para la determinación de las propiedades de las remotas y antiguas galaxias primitivas están presentes en nuestra galaxia. Para calcular el espectro de una galaxia primitiva, estimamos simplemente la abundancia de los diversos objetos en tal galaxia, tomamos sus espectros en la forma en que aparecen en nuestra galaxia y sumamos los espectros de acuerdo con la abundancia e importancia de los objetos.

La tarea de determinar el grado de predominio de los objetos en una galaxia primitiva queda simplificado por el hecho de que la mayor parte de los objetos son descendientes de las masivas estrellas que de manera continua se forman y mueren. (Las estrellas de pequeña masa contribuyen solamente con la luz visible y viven demasiado tiempo para crear ningún objeto interesante.) Además, los objetos descendientes tienen una vida tan corta (comparada con el tiempo de colapso de la protogalaxia) como las propias estrellas masivas, de modo que el número de esos objetos que contribuye al espectro de la galaxia primitiva es proporcional a la velocidad de formación de estrellas. La velocidad de formación de estrellas en una galaxia primitiva brillante puede ser 3000 veces mayor que la actualmente existente en nuestra galaxia, de modo que en una galaxia primitiva es de esperar un número 3000 veces mayor de estrellas masivas y sus descendientes de vida corta.

Para determinar las clases de tales descendientes, podemos seguir la evolución de una estrella masiva desde su nacimiento hasta su muerte. Se cree que la estrella se forma dentro de una nube fría (10 grados Kelvin) consistente principalmente en hidrógeno molecular (H_2), monóxido de carbono (CO) y granos de polvo de silicatos y carbono. El hidrógeno molecular no emite mucha radiación, de modo que las únicas ondas importantes que emanan de la nube son las asociadas a la línea espectral, en el dominio de las microondas, del monóxido de carbono (a una frecuencia de 115 gigahertz). La estrella luce brillantemente en las regiones visible y ultravioleta del espectro, aunque tal radiación es absorbida por la nube y, por tanto, no puede observarse fuera de ésta. La radiación



ESPECTRO HIPOTETICO DE UNA GALAXIA PRIMITIVA (*arriba*), **obtenido sumando los espectros de los fenómenos contribuyentes, tales como los que figuran en las dos ilustraciones precedentes. Este espectro es el intrínseco; el observado aparecería trasladado en frecuencia hacia la izquierda según el corrimiento hacia el rojo preciso. La mayor incertidumbre en la determinación del espectro proviene de la estimación del cociente del número de estrellas masivas que están envueltas por polvo al número de que no lo están. Aquí se ha supuesto que la relación es de 1:1. Las regiones visible y ultravioleta** (*color*), **que se muestran a mayor escala en la parte inferior, comprenden líneas de emisión de hidrógeno (las líneas de Balmer** $H\alpha$ **y** $H\beta$ **y la línea de Lyman** $L\alpha$), **carbono** (C^+ **y** C^{++}), **oxígeno** (O^+ **y** O^{++}) **y magnesio** (Mg^+) **y líneas de absorción de carbono** (C^{+3}) **silicio** (Si^{+3}), **oxígeno** (O **y** O^{+5}), **nitrógeno** (N^{+4}) **e hidrógeno (la línea Lyman** $L\beta$). **Las líneas han sido ensanchadas por los corrimientos Doppler de gas moviéndose a velocidades de algunos cientos de kilómetros por segundo. (Dibujo de Gabor Kiss.)**



ESPECTRO DE UN QUASAR (*arriba*). Es muy diferente del hipotético espectro de una galaxia primitiva. Ambos objetos pueden distinguirse aun cuando pudieran presentar el mismo aspecto a través del telescopio. El espectro de un quasar no tiene ni las emisiones continuas ni las líneas de emisión que son de esperar en la galaxia primitiva procedentes de estrellas calientes y gas. El espectro de una galaxia primitiva presenta un máximo en el infrarrojo y otro en la región ultravioleta, mientras que el espectro de un quasar presenta un descenso continuo y suave. Las líneas de las regiones visible y ultravioleta (*color*), que se muestran a mayor escala abajo, son aproximadamente cien veces más anchas. Este espectro intrínseco es el del quasar 3C 273, uno de los primeros descubiertos. El espectro observado está trasladado hacia la izquierda en un factor de 1 más el corrimiento hacia el rojo 0,158.

calienta los granos de polvo que envuelven la nube hasta unos 30 grados Kelvin, temperatura a la cual los granos reemiten la energía de la estrella en forma de emisión térmica en el infrarrojo. También se forma una espesa capa de gas ionizado alrededor de la estrella recubierta. Aunque la radiación visible y ultravioleta emitida por la estrella y el gas no puede penetrar a través de los granos de polvo, las radioondas sí pueden hacerlo.

Pronto la capa de gas ionizado se dilata, descubriendo la estrella por haber dispersado o destruido la mayor parte de los granos de polvo. Como resultado, la intensidad de la radiación infrarroja decrece bruscamente y la intensidad de la radiación visible crece. (La nebulosa de Orión en nuestra galaxia es un ejemplo de una región de estrellas calientes visibles y gas ionizado). Cuando la estrella masiva alcanza el final de su ciclo de vida, explota violentamente en forma de supernova. Una explosión de supernova debe ser un suceso espectacular en una galaxia normal a causa de la comparativa palidez de las otras estrellas. En una galaxia primitiva, sin embargo, una supernova pasaría desapercibida a las longitudes de onda del visible, a causa de la abundancia de otras estrellas extremadamente brillantes y masivas.

La situación cambia en otras longitudes de onda. El compacto resto de la supernova, que podría ser un pulsar o un agujero negro, y la capa de gas en expansión que rodea a dicho resto, aumentarían en forma apreciable las emisiones de la galaxia en las bandas de radio y rayos X. Si el resto compacto es una estrella de neutrones pulsante, emitiría intensamente radioondas y rayos X. Podría también aportar rayos X en una forma distinta y desusada. Si, como ocurre con frecuencia, la estrella original era parte de un sistema binario, con una estrella cuya duración de vida es mayor, la estrella de neutrones puede quedar en órbita alrededor de la estrella de vida más larga. Se podría entonces producir transporte de masa a la estrella de neutrones mediante un "viento" estelar o por gravedad, de modo que la estrella de neutrones atraiga materia arrancada de la otra estrella. El resultado sería un sistema binario emisor de rayos X como los que se han encontrado en nuestra galaxia. Una descripción semejante se aplicaría si, en lugar de una estrella de neutrones, el resto binario fuese un agujero negro.

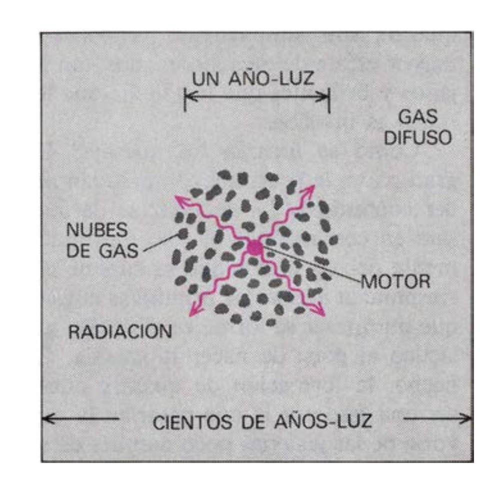
Los pulsares y los restos de supernova probablemente contribuyen también al

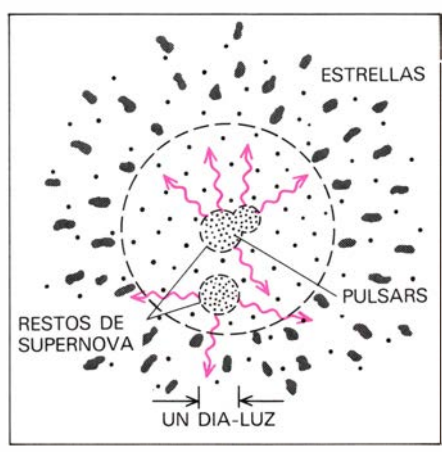
espectro dando origen a rayos cósmicos. El campo magnético del pulsar, en rápida rotación, y la fuerte onda de choque del resto de la supernova, son capaces de acelerar los electrones, protones y otras partículas subatómicas hasta velocidades próximas a la de la luz. A medida que esas partículas se distancian del pulsar o del resto de supernova e interaccionan con los campos magnéticos de la galaxia, emiten luz y otras formas de radiación electromagnética. Cuando encuentran átomos del medio interestelar, pueden generar rayos gamma.

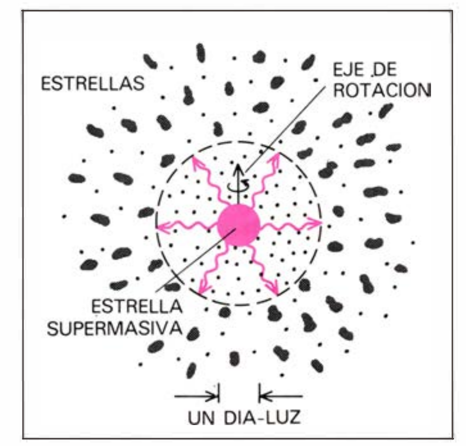
El espectro de una galaxia primitiva puede obtenerse ahora sumando los espectros de los fenómenos participantes. La mayor incertidumbre al estimar el espectro proviene de la estima de la relación de estrellas masivas que quedan envueltas por polvo al de aquellas que no lo están. Hemos supuesto que la relación es 1:1, lo que debe ser suficiente a menos que una situación predomine abiertamente. Para corrimientos al rojo de entre 3 y 30, el espectro calculado puede distinguirse con extrema facilidad con telescopios ópticos o de infrarrojo próximo, como el telescopio espacial. Otros instrumentos que están todavía en fase de proyecto, tales como los observatorios orbitales del infrarrojo lejano y de rayos X, deben ser capaces de detectar distantes galaxias primitivas a otras longitudes de onda.

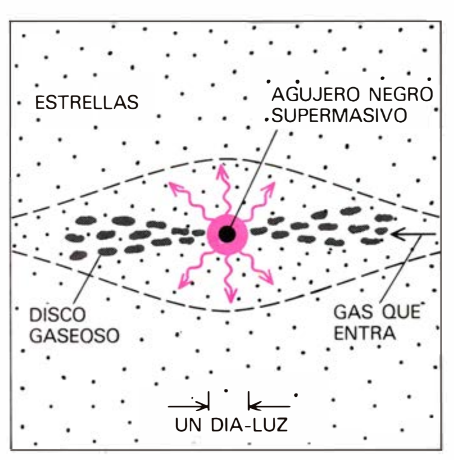
¿Hay algunos detalles en esos espectros que diferencien una galaxia primitiva de los otros objetos? En el espectro visible y ultravioleta de una galaxia primitiva es de esperar un espectro continuo de emisión procedente de las estrellas, un espectro de líneas de absorción de la atmósfera de las estrellas y un espectro de líneas de emisión del gas ionizado que rodea las estrellas calientes. El gas y las estrellas en una galaxia primitiva se están moviendo rápidamente y de forma caótica, de modo que las líneas serán mucho más anchas que las del gas y las estrellas, de movimiento más lento, que no sufrirán corrimientos Doppler tan importantes respecto a sus posiciones normales. Las velocidades correspondientes a las anchuras de línea no deben exceder la velocidad de escape de una galaxia primitiva: algunos cientos de kilómetros por segundo. Un objeto que presente tales características ha de ser un firme candidato para su clasificación como galaxia primitiva.

Cuando se descubrieron los quasars, se pensó que podría tratarse de galaxias primitivas; y ello, sobre todo, en virtud de sus grandes corrimientos hacia el rojo. Sin embargo, el espectro de un quasar es muy diferente del calculado para una galaxia primitiva. El espectro continuo de 3C 273, uno de los primeros quasars identificados, presenta un suave descenso desde el infrarrojo hasta el ultravioleta, mientras que una galaxia primitiva debería tener un máximo en la región infrarroja y otro en la visible y ultravioleta. Además, el espectro continuo de un quasar está frecuentemente polarizado: las ondas luminosas oscilan en una dirección preferida. Por el contrario, la radiación de una galaxia primitiva se espera que presente poca o ninguna polarización. El espectro de un quasar tiene también anchuras de línea mucho más grandes, lo que sugiere que tal objeto es un fenómeno más energético que una galaxia primitiva. El gas de los quasars parece estar moviéndose a velocidades del orden de 10.000 kilómetros por segundo. Si las nubes de una galaxia primitiva viajaran a tales velocidades, habrían escapado del campo gravitatorio de la galaxia hace mucho tiempo, a menos que se hubieran visto implicadas en una reciente y violenta explosión o que estuvieran ligadas a una masa grande y compacta, tal como una estrella supermasiva o un agujero negro.









"MOTOR" DE UN QUASAR, situado, probablemente, en el centro de un agrupamiento de densas nubes inmersas en gas difuso (*parte superior izquierda*). Se han propuesto tres modelos teóricos del motor. En el primero (*parte superior derecha*), el motor es un compacto cúmulo de estrellas. En el segundo modelo (*parte inferior izquierda*) es una estrella supermasiva o un pulsar. En el tercer modelo (*abajo derecha*) es un agujero negro supermasivo. Todo parece indicar que no son galaxias primitivas.

Los quasars pueden distinguirse también de las galaxias porque el brillo de su luz varía en el curso de sólo unos pocos meses o años. La luz de los objetos BL Lacertae, entidades semejantes a quasars con peculiaridades espectrales propias, varía de intensidad en el curso de sólo unas horas. Tales variaciones son imposibles para un conjunto de 100.000 millones de estrellas, que no pueden variar al unísono. Aun la explosión de supernovas, como hemos indicado arriba, apenas puede alterar el brillo de una galaxia primitiva, y ciertamente no puede causar una variación en el curso de horas. Los quasars variables deben ser extremadamente compactos. Son, con probabilidad, menores que la distancia que la luz puede recorrer en el tiempo necesario para que la intensidad de su luz varíe.

Si los quasars no son galaxias primitivas, ¿qué son? Para responder a esta pregunta, los astrónomos han desarrollado modelos de la región que se cree emite las líneas presentes en el espectro

del quasar. En un modelo típico, el quasar está envuelto por una capa de gas de baja densidad que se extiende a cientos de años-luz y tiene líneas espectrales con una anchura que corresponde a una velocidad de 1000 kilómetros por segundo. Dentro de esta región, hay un compacto conjunto de nubes gaseosas densas y en rápido movimiento, cada una de una dimensión aproximada de un año-luz. El grupo de nubes constituye la fuente de las líneas más anchas en el espectro del quasar. En el centro del quasar está el "motor", de tamaño a lo más de un día-luz, que emite el espectro continuo e ioniza las nubes más densas.

La naturaleza exacta del motor del quasar es objeto de activa investigación en la actualidad. Se han propuesto tres modelos hasta ahora. En el primero, el motor es un compacto cúmulo de estrellas cuya energía proviene de una multitud de choques estelares, explosiones de supernova y pulsars que han sido comprimidos en una región de radio quizá sólo diez veces mayor que el radio del sistema solar. En el segundo modelo, el motor es una estrella supermasiva o un pulsar, cuya energía proviene de la contracción gravitatoria o rotación de una estrella magnética única con una masa enorme, del orden de mil millones de masas solares. En el tercer modelo, análogo al modelo actual de las fuentes binarias de rayos X, tales como el Cisne X-1, el motor es un agujero negro supermasivo de mil millones de masas solares. La energía no procede del propio agujero negro, sino del gas y las estrellas descompuestas que chocan entre sí al precipitarse hacia él. En los choques, el gas y las estrellas liberan aproximadamente una décima parte de la energía que corresponde a su masa en reposo (su masa en reposo multiplicada por el cuadrado de la velocidad de la luz). Una pérdida de sólo diez masas solares por año suministrada al agujero negro aporta energía suficiente para alimentar el quasar más brillante y para superar claramente el brillo de una galaxia de un billón (un millón de millones) de estrellas.

¿Cuál es el origen de los quasars? Los modelos del motor del quasar sugieren que éste existe en el núcleo de galaxias que son por lo demás normales, donde los compactos cúmulos de estrellas, estrellas supermasivas y grandes agujeros negros tienen mayor facilidad para formarse. Los objetos BL Lacertae, extensas radiofuentes que semejan radio-quasars, y los núcleos Seyfert, todos los cuales son menos luminosos que los quasars pero tienen virtualmente el mismo espectro, han sido hallados en galaxias. Es razonable suponer que los quasars son simplemente versiones a mayor escala de esos fenómenos, tan lejanos y brillantes que la galaxia que los rodea es invisible.

¿Cómo se forman los quasars? En gran parte, la pregunta continúa sin poder contestarse. La abundancia de quasars en corrimientos al rojo inmediatamente debajo de los que se supone corresponden a galaxias primitivas sugiere que un quasar se forma en el núcleo galáctico al poco de nacer la galaxia. De hecho, la formación de quasars puede ser una fase por la que pasarían la mayoría de las galaxias poco después de su nacimiento. En el modelo del colapso de una galaxia descrito más arriba, no es preciso que la actividad se detenga cuando cesa la formación de estrellas. Pueden producirse sucesos que obliguen a la joven galaxia a desarrollar en su centro la clase de objeto supermasivo requerido por los modelos del motor del quasar. (Se han indicado como procesos posibles el choque y soldadura de estrellas comprimidas en el núcleo galáctico o la unión y colapso de nubes gaseosas residuo de la formación de estrellas. Otra posibilidad es que se forme lentamente un gran agujero negro a partir de otro pequeño que va devorando estrellas vecinas.) Entonces, mediante un mecanismo mal conocido, el objeto supermasivo brilla durante un corto tiempo mucho más intensamente de lo que nunca lo hizo la galaxia. Por último, el quasar muere, dejando visible la galaxia circundante. En el momento presente, este modelo es poco más que una especulación. Los astrónomos esperan descubrir, a lo largo de los próximos años, más indicaciones acerca de la manera como están relacionadas la formación de quasars y la formación de galaxias.

Todas las indicaciones sugieren que los quasars no son galaxias primitivas. Sin embargo, algunas galaxias primitivas podrían haber sido tomadas por quasars. Antes de explorar detenidamente el cielo en busca de nuevos candidatos para galaxias primitivas, será necesario examinar la lista de los quasars ya catalogados en busca de aquellos que tienen líneas de emisión cuya anchura corresponde a velocidades de sólo unos pocos cientos de kilómetros por segundo. Tal examen puede ser un proyecto hercúleo, porque ya se han catalogado cientos de quasars. Además, los espectros de muchos de esos quasars no se han registrado con suficiente detalle. Sin embargo, el examen se podría llevar a cabo si se estudiaran primero los quasars con máximo corrimiento hacia el rojo. El próximo paso será obtener nuevos candidatos haciendo uso de la técnica del prisma-objetivo. Esta técnica registra los espectros aproximados de muchos objetos en el cielo al mismo tiempo, permitiendo estimas rápidas del corrimiento hacia el rojo, anchuras de bandas y otras características. Ya ha mostrado su utilidad en la búsqueda de quasars. Un prisma objetivo empleado con el telescopio espacial será quizás el sistema más indicado para detectar una galaxia primitiva. A través del telescopio, los objetos muy compactos aparecerán como puntos, mientras que las galaxias primitivas se verán con cierto detalle. Esto puede eliminar la necesidad de obtener espectros detallados para distinguir entre las galaxias primitivas y los quasars.

El descubrimiento de una galaxia primitiva constituiría un triunfo importante para la astronomía moderna. Las actuales teorías de cosmología, formación de galaxias y evolución estelar tendrían nuevas pruebas que constituirían fuertes argumentos en su favor. En un campo tan complejo como la astronomía, sin embargo, las predicciones teóricas no siempre son exactas; es perfectamente posible que fenómenos imprevistos pudieran complicar la búsqueda de galaxias primitivas. Por ejemplo, podría suceder que no todas las galaxias se hubiesen formado en el violento colapso de una extensión de gas y polvo con una masa de unas 10^{11} masas solares. Quizás algunas galaxias se formaron de la reunión de cúmulos de estrellas mucho menores, a su vez creados a partir de las numerosas perturbaciones de pequeña masa (desde 10^6 hasta 10^{11} masas solares) existentes en el universo recién nacido. Una galaxia primitiva formada por reunión sería, presumiblemente, más débilmente luminosa que una formada por el proceso de colapso que hemos descrito, ya que la formación de estrellas sería más lenta.

La búsqueda se podría complicar todavía más si el polvo fuera tan abundante en las galaxias jóvenes que envolviera a la mayor parte de las estrellas masivas. En el peor de los casos, la totalidad de la galaxia podría quedar oculta a la vista. Transcurrirán varios años antes de que se introduzcan en el espacio instrumentos capaces incluso de detectar la radiación infrarroja emitida por tal polvo, y pasarán muchos años antes de que el corrimiento hacia el rojo de una galaxia cubierta de polvo se pueda medir de manera fidedigna. La mayor parte de los astrónomos, sin embargo, creen que algún día se encontrarán las galaxias primitivas. Entonces se hará posible estudiar directamente la formación de galaxias.

Ecología de los escarabajos estercoleros africanos

Los escarabajos estercoleros desempeñan un papel ecológico clave al eliminar los excrementos que dejan los rebaños de mamíferos. El éxito en la competencia que se establece entre estos escarabajos depende de su temperatura corporal

Bernd Heinrich y George A. Bartholomew

A cualquiera que contemple los grandes rebaños de las praderas y sabanas del Africa oriental se le puede ocurrir que los animales producen grandes cantidades de heces. Pero sobre el suelo no se ve mucho estiércol. ¿Dónde ha ido a parar? La mayor parte del estiércol es eliminado y enterrado rápidamente por legiones de escarabajos de la familia Escarabeidos, los llamados escarabajos estercoleros, para los cuales constituye un recurso vital. Los escarabajos hacen mucho más que eliminar un material que de otro modo se acumularía en el suelo, sofocando las plantas y limitando así probablemente las poblaciones de animales que la tierra puede mantener. Su actividad, además, fertiliza y airea el suelo, retarda la expansión de parásitos y organismos productores de enfermedades y reduce el número de moscas molestas que crían en el estiércol. La vida de estos escarabajos ilustra las intrincadas relaciones que se dan en un ecosistema, y revela asimismo algunas adaptaciones fisiológicas y de comportamiento, muy interesantes, de estos insectos.

Los escarabajos estercoleros, de los que sólo en Africa existen más de 2000 especies, han evolucionado hasta alcanzar tamaños y formas muy diversos. Los mayores pesan más de 20 gramos y sobrepasan, en siete u ocho veces, el tamaño de los pájaros, murciélagos y musarañas más pequeños. Los escarabajos situados en el polo opuesto de la gama miden mil veces menos y apenas si pesan unos miligramos. En circunstancias apropiadas, estos insectos son tan abundantes que el estiércol resulta un recurso escaso, aunque el suministro es constante. De hecho, su disponibilidad puede limitar la reproducción y el crecimiento de los escarabajos. En Africa tropical muchas especies compiten fuertemente por el estiércol, y se ha desarrollado una notable variedad de pautas de utilización del mismo.

Fuimos a estudiar los escarabajos estercoleros al Parque Nacional Tsavo, en Kenia, que posee la mayor población de elefantes del mundo. Un grupo de cuatro o cinco elefantes puede procesar una tonelada métrica de alimentos al día, gran parte del cual vuelve al suelo en forma de estiércol. En el Parque Tsavo, durante la estación de las lluvias, la eliminación de las heces de elefante por parte de los escarabajos es asombrosamente rápida. Durante el día son pocos los escarabajos atraídos por las heces frescas de los elefantes, pero después de la puesta del sol estos insectos llegan volando en grandes nubes zumbadoras. Pequeñas muestras de boñigas de elefante que recolectamos durante el día y colocamos como cebo durante la noche atrajeron verdaderos enjambres de coleópteros. Más de 3800 acudieron a una muestra de medio litro que estuvo expuesta sólo durante 15 minutos. Durante la primera parte de este período aparecieron únicamente unos pocos escarabajos, pero después la afluencia no cesó.

Otra muestra de unos 30 litros de heces de elefantes desencadenó una pavorosa y súbita acometida. Sobre el estiércol se posaron nubes de pequeños escarabajos que inmediatamente empezaron a excavar en él. Al cabo de una hora y media el montón de estiércol se había transformado en una palpitante alfombra en expansión, constituida por una capa fluidificada de escarabajos y heces húmedas cubierta por una delgada lámina de material fibroso: los restos gruesos y no digeribles de las plantas. Los escarabajos transformaron los duros bolos de estiércol, malolientes y del tamaño de un balón de fútbol, en un felpudo aplastado de dos o tres centímetros de grosor y de hasta dos metros de diámetro.

A un observador no le será difícil apreciar que la competencia intra– e interespecífica por el estiércol de elefante es intensa entre las muchas especies que son atraídas por aquél. ¿Cuáles son las consecuencias? ¿Cómo se han acomodado a ello los escarabajos estercoleros? Nos ocuparemos de estas cuestiones al describir algunas de las pautas que exhiben los escarabeidos al procurarse y utilizar el estiércol y, más tarde, al examinar algunas de las respuestas fisiológicas y de comportamiento que operan dentro de estas pautas.

Unas cuantas especies de escarabajos se especializan en los excrementos de determinados mamíferos, pero la mayoría acepta cualquier tipo de estiércol que encuentre. Pueden distinguirse tres pautas principales en la utilización del estiércol. Algunos escarabajos, denominados Endocópridos, excavan en el estiércol y permanecen allí viviendo y alimentándose hasta que éste se termina o su estructura se descompone. Otras especies perforan la tierra situada debajo o cerca de un montón de estiércol y acarrean los excrementos al interior de la madriguera. Y los hay que, como el escarabajo sagrado, *Scarabaeus sacer*, cortan fragmentos de estiércol, los moldean formando una pelota y los arrastran de uno a 15 metros, o incluso más, antes de enterrarlos.

Los escarabajos que utilizan cada una de las tres técnicas tienen adaptaciones morfológicas que complementan su comportamiento. Los Endocópridos son generalmente pequeños, característica adecuada para comer en el interior de excrementos que contienen mucho material fibroso, como los del elefante y los rinocerontes. Los escarabajos minadores suelen ser grandes y robustos; su cuerpo funciona como una máquina excava-



ESCARABAJOS ESTERCOLEROS del Africa oriental, ocupados en eliminar un montón de heces de elefante durante la noche. Las manchas brillantes son el reflejo del dorso de los escarabajos de los que pueden contarse más de 200; los individuos visibles pesan en conjunto más que el estiércol.

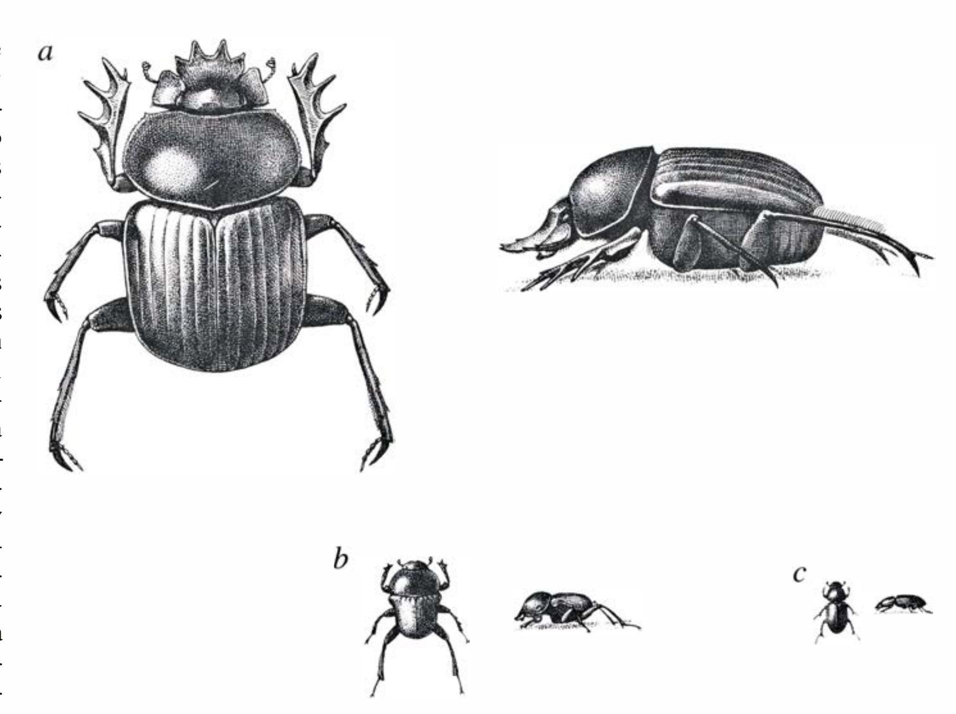


MOLDEADO DE UNA PELOTA a partir de un bolo de excremento de elefante que realiza un escarabajo pelotero de la especie *Kheper platynotus*, activo durante el día (la mayoría de escarabajos estercoleros de Africa oriental trabajan durante la noche). A la izquierda de la ilustración se aprecian dos pelotas ya terminadas, con un escarabajo que trepa a una de ellas. *Kheper platynotus* es la especie que se ha ilustrado en la portada de este número. El pelotero macho empuja la pelota terminada, con la hembra encaramada sobre la misma, a lo largo de una distancia de varios metros hasta que encuentra un lugar adecuado para enterrarla. El suelo laterítico de la región de Africa oriental debe su color rojo a los óxidos de hierro.



ALFOMBRA DE RESIDUOS, que es todo lo que quedó de un bolo de excrementos de elefante después de que un gran número de escarabajos hubiera actuado sobre él durante la noche subsiguiente a su deposición. La mayoría de los coleópteros que se encontraban allí durante la noche habían desaparecido en el momento de tomarse esta fotografía, aunque los que permanecían todavía bajo la alfombra se contaban por millares; gran número de los mismos desaparecería al día siguiente. Enterrar el estiércol tiene la ventaja de sustraerlo del posible alcance de las otras especies competidoras.

dora. El ejemplo más sorprendente de esta adaptación está representado por los grandes escarabajos de la especie *Heliocopris dilloni*, que se alimentan sólo de estiércol de elefante. Algunos de estos escarabajos pueden pesar hasta 25 gramos. Tienen patas vigorosas, con las tibias anteriores espatuladas, y espinas dirigidas hacia atrás sobre las tibias de sus dos pares de patas posteriores, que les permiten obtener una buena tracción mientras se abren camino a través del suelo, avanzando con la cabeza. Las superficies anterior y superior de la cabeza recuerdan la pala rasadora de un *bulldozer*. Los escarabajos peloteros, en cambio, tienen las patas posteriores largas y delgadas, arqueadas hacia afuera, adecuadas para correr y que pueden doblarse alrededor de una masa de estiércol, lo que les permite sujetarse a ella mientras la moldean formando una pelota. Sus tibias anteriores tienen extensiones en forma de rastrillo que utilizan para ir añadiendo pedacitos de estiércol a la pelota mientras ésta va siendo moldeada. Los escarabajos peloteros tienen tamaños muy diversos. Algunos pesan sólo unos pocos miligramos y hacen rodar pelotas del tamaño de un guisante. Otros, como los de las especies *Scarabaeus laevistriatus* y *Kheper platynotus*, pesan hasta 10 gramos y de las boñigas de elefante cortan esferas del tamaño de una pelota de tenis.



TRES ESPECIES de escarabajos estercoleros de Africa oriental, dibujadas en visión dorsal y lateral. La escala es la misma para todos ellos. El individuo mayor (que no es el escarabajo estercolero más grande de Africa) es *Scarabaeus laevistriatus*, una especie de escarabajo pelotero activa al ocaso y durante la noche. Abajo a la izquierda hay un escarabajo pelotero nocturno más pequeño, *Gymnopleurus laevicollis*. El tercer ejemplar (*c*) es un endocóprido; este animal excava en los excrementos y se alimenta de ellos hasta que los recursos se han agotado. Los Endocópridos suelen ser de talla pequeña.

Enterrar el estiércol tiene la ventaja de sustraerlo del alcance de los competidores. Los escarabajos que entierran las pelotas de estiércol suelen verse obligados a trabajar rápidamente. Las especies que entierran el estiércol directamente bajo el acúmulo de heces (sin hacerlo rodar) suelen formar parejas sexuales y trabajan en equipo. Acarrean el estiércol hacia el interior de la madriguera en montoncitos sueltos, terminando por moldear parte del mismo en pelotas de incubación redondas o piriformes. En cada pelota se deposita un solo huevo, y así la larva tendrá a su disposición todo el alimento y el agua que necesite para completar su desarrollo. Unicamente las hembras hacen pelotas de incubación; y ellas se ocupan también de casi todas las tareas relacionadas con la anidación.

Por lo general, la hembra hace la madriguera inicial, arrastrando a la superficie la tierra que remueve y dejándola allí; el macho aparta entonces la tierra de la entrada. Tan pronto como la hembra deja de aportar tierra, el macho toma un fragmento de estiércol con sus patas anteriores y penetra en el túnel retrocediendo. Cuando el extremo de su abdomen entra en contacto con la cabeza de la hembra, abandona su carga y vuelve a la superficie a por más. La hembra toma el fragmento de estiércol y lo transporta por el túnel el trecho que falta. En esta pauta de comportamiento el macho trabaja en, o cerca de, la superficie; se halla expuesto, pues, a los depredadores, en tanto que la hembra, que contribuye en mayor medida al esfuerzo reproductor, trabaja en la relativa seguridad de una madriguera que puede hundirse un metro o más bajo la superficie.

Tanto los machos como las hembras de los escarabajos peloteros hacen pelotas de estiércol. Ambos se comen algunas pelotas, pero otras (hechas por el macho) son elementos clave en la secuencia de apareamiento. El macho y la hembra se encuentran por primera vez en el lugar y en el momento de la caída del excremento, que los ha atraído a ambos. La hembra puede ayudar al macho que esté haciendo una pelota o que ya la haya terminado. Los detalles varían de una especie a otra, pero la pauta general del comportamiento reproductor es la misma.

Kheper platynotus, un gran escarabajo que actúa durante el día, puede servir de ejemplo. La hembra suele unirse al macho en el montón de estiércol, mientras aquel va formando la pelota. Cuando el macho empieza a rodar la bola terminada, la hembra se sujeta a la misma y es transportada sobre ella; aquél la empuja trabajosamente hasta el lugar del enterramiento. El macho entierra la pelota excavando la tierra bajo ella, de suerte que la bola de estiércol se hunde verticalmente en el suelo. La hembra, todavía a bordo, cae en el hoyo con la pelota. Cuando ésta ha sido enterrada, la pareja se alimenta de ella, se aparea y vuelve a la superficie.

En algunas especies la hembra no cabalga sobre la pelota. Por ejemplo, la hembra de la especie *Gymnopleurus laevicollis* se mantiene erguida en el suelo sobre sus patas traseras, poniendo las delanteras sobre el lado anterior de la pelota. Da la impresión de que estuviera tirando de la bola hacia sí, mientras el macho, en el otro lado, se yergue sobre sus patas delanteras y empuja con las posteriores, como es usual en los escarabajos peloteros. La hembra de la especie *Scarabaeus sacer* sigue al macho, a unos dos o tres centímetros de distancia, cuando éste hace rodar la pelota hasta el lugar del enterramiento.

En todas las especies de escarabajo pelotero el macho realiza el trabajo de enterrar la bola y la hembra la mordisquea durante el apareamiento. Debido a que la pelota de estiércol sirve, en cierto modo, de presente que el macho ofrece a la hembra y que le permite aparearse, se la ha denominado pelota nupcial. (La alimentación nupcial se da también en otros insectos y en muchos vertebrados.) Resulta evidente que el éxito en la competencia por el alimento se halla estrechamente ligado al éxito reproductor.

Aunque el escarabajo pelotero macho

no corre peligro de ser comido por la hembra después de la cópula, que es lo que ocurre en algunos insectos como los Mántidos, sí se expone a un riesgo considerable al moldear la pelota nupcial. Los acúmulos de estiércol atestados de escarabajos atraen a aves (calaos, faisanes de espuelas, pintadas) y mamíferos (mangostas) que depredan a estos coleópteros. Al construir una pelota de estiércol para que la hembra la coma bajo tierra, en una relativa seguridad, el macho se está poniendo a sí mismo en peligro, al tiempo que reduce el riesgo para su pareja en potencia.

Entre los escarabajos peloteros de Africa oriental es intensa la presión selectiva a favor de una aceleración en el proceso de localización del estiércol y de su extracción de los montones de excrementos. La presión existe no sólo porque el estiércol atrae a depredadores de escarabajos, sino también porque el ambiente físico y la competencia entre los coleópteros obligan a que el escarabajo se apresure tanto en llegar como en marcharse. Nada hay pues de extraño en que uno se pregunte qué mecanismos fisiológicos y de comportamiento se han desarrollado para la aceleración del proceso de obtención del estiércol.

Los escarabajos estercoleros no pueden emplear de forma ordinaria la estrategia obvia de seguir a los rebaños de mamíferos, porque éstos son activos tanto durante el día como durante la noche, mientras que los escarabajos (según las especies) son diurnos o nocturnos. Además, los escarabajos que han enterrado una pelota permanecen con ella por lo menos durante un día, al final del cual es probable que los mamíferos que depositaron los excrementos estén ya muy lejos.

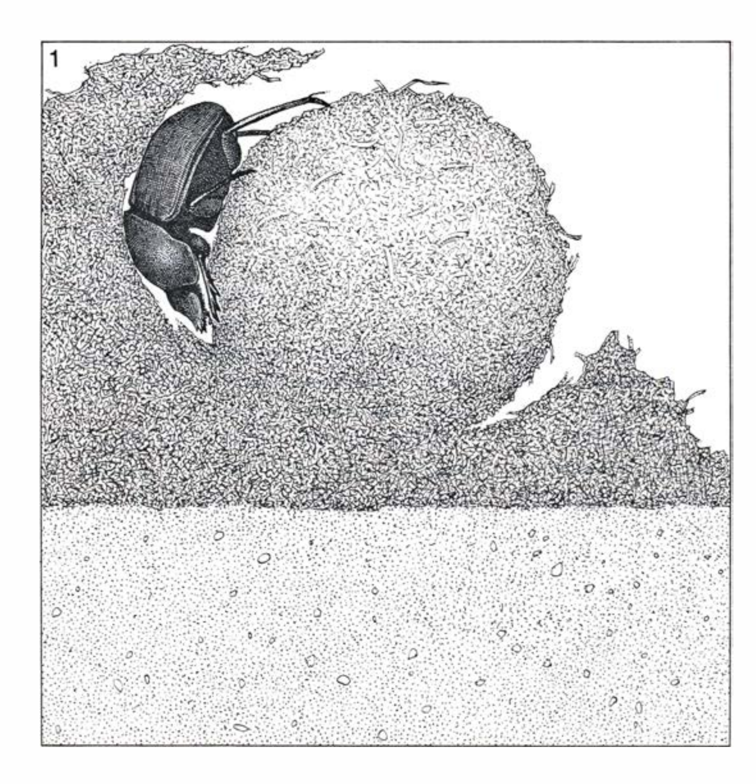
Resulta evidente que los coleópteros encuentran los excrementos frescos por el olor: siempre se acercan en contra de la dirección del viento. Descubrimos que cantidades enormes de escarabajos lograban encontrar cebos de estiércol que habían sido escondidos, aunque se tratara de zonas sin animales ungulados.

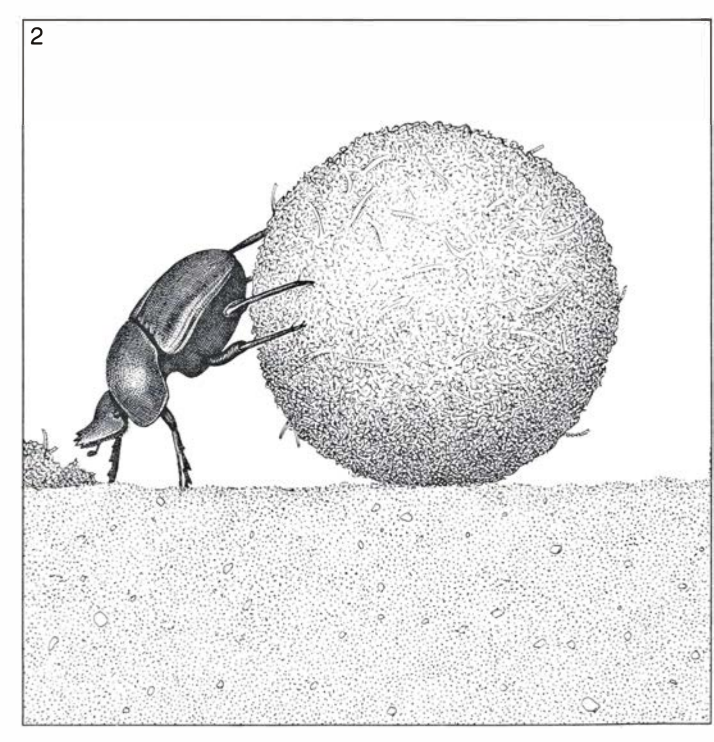
Se ignora si los escarabajos vuelan continuamente en busca de estiércol o bien esperan a emprender el vuelo hasta percibir el olor. Se sabe, sin embargo, que el control de la temperatura corporal desempeña un importante papel en la velocidad, el gasto energético y la potencial recompensa de hallar estiércol fresco. La temperatura corporal de un escarabeido en reposo no suele diferir en más de un grado Celsius de la temperatura ambiental. Algunos insectos de gran tamaño tienen durante el vuelo una temperatura torácica de hasta 35 grados C por encima de la temperatura del aire. En cualquier especie de insectos, la potencia de los batimientos del ala y la velocidad del vuelo se hallan estrechamente relacionadas con la temperatura muscular. Si la temperatura de los músculos del vuelo en el tórax es inferior a unos 34 grados, muchos insectos grandes no pueden generar las fuerzas ascensional y propulsora necesarias para permanecer en vuelo. ¿Qué ocurre en los escarabajos estercoleros de distintos tamaños?

Medimos la temperatura torácica de escarabajos en vuelo capturándolos durante la noche cuando llegaban al estiércol o a la luz e insertando inmediatamente en sus músculos torácicos una delgada aguja hipodérmica que portaba un diminuto termopar. La temperatura torácica de los escarabajos grandes era de 39 a 45 grados, lo que representaba una temperatura de dos a ocho grados por encima de la de la mayoría de mamíferos, y tan alta como la de las aves voladoras. Algunos escarabajos habían estado volando con temperaturas musculares de sólo uno o dos grados por debajo del nivel que es perjudicial y posiblemente letal.

Estas elevadas temperaturas torácicas son necesarias para el vuelo en los mayores escarabajos estercoleros; no pueden permanecer en vuelo con temperaturas musculares bajas. Sin embargo, los escarabeidos más pequeños pueden volar bien con una temperatura corporal de sólo 27 grados, que supone únicamente un par de grados por encima de la temperatura ambiente nocturna del Parque Tsavo en la época en que estuvimos trabajando allí.

¿Qué es lo que explica esta diferencia? Por lo menos en parte, la respuesta tiene





PROCESO DE ELABORACION de una pelota de estiércol por una hembra de la especie *Kheper aegyptiorum*. En primer lugar el animal corta el material para la pelota de un bolo de estiércol y después le da forma (*1*). Cuando el ovillo está terminado, la hembra empieza a hacerlo rodar hacia un lugar

que ver con la relación entre tamaño y tasa de flujo térmico. A partir de nuestras extensivas mediciones de la temperatura torácica en función del tamaño, llegamos a la conclusión de que, en los escarabajos que pesan hasta dos gramos, la temperatura corporal es una función pasiva de la producción de calor (una consecuencia necesaria del metabolismo de vuelo) y de la tasa de enfriamiento pasivo. Cuanto mayor es el escarabajo, tanto menor es la cantidad relativa de superficie de disipación del calor y tanto mayor es la temperatura corporal. Existe, no obstante, un límite superior de temperatura tolerable, unos 45 grados, temperatura a la que de hecho se acercan aquellos escarabajos que pesan sólo dos o tres gramos.

Los escarabajos mayores deben disipar activamente parte del calor que generan en el intenso metabolismo asociado con el vuelo. Sólo pueden hacerlo cuando el gradiente de temperatura entre su cuerpo y el aire es grande. Puesto que, inevitablemente, se calientan, puede deducirse que han desarrollado los mecanismos bioquímicos para que el motor de vuelo funcione de manera óptima a la temperatura que produce durante el vuelo. Los escarabajos pequeños no se calientan; han sufrido, pues, presión selectiva para volar a temperaturas musculares menores.

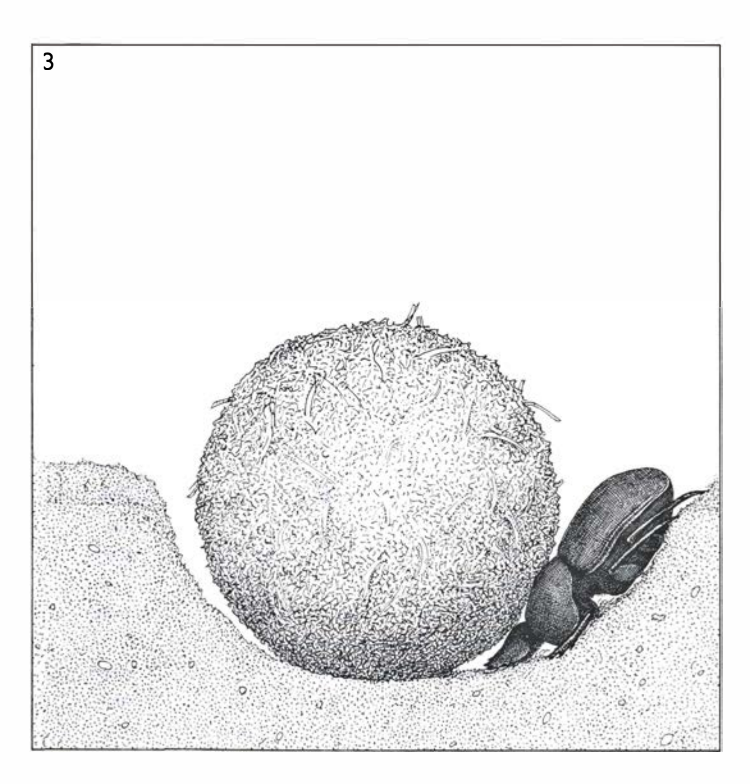
Cuando se ha obligado a los enzimas y a otros componentes de la maquinaria bioquímica a funcionar a velocidades máximas a una temperatura determinada, su rendimiento por lo general disminuye a otras temperaturas. Esta situación compensadora puede tener implicaciones significativas en las estrategias de competencia por el estiércol. Los escarabajos pequeños pueden levantar el vuelo inmediatamente cuando huelen los excrementos; en cambio, los escarabajos grandes deben gastar tiempo y energía en caldearse antes de remontar el vuelo. (Comprobamos que un escarabajo de la especie *Heliocopris dilloni* que pesaba 11,7 gramos necesitaba cinco minutos para calentarse desde 27 grados hasta una temperatura de vuelo de 40 grados.)

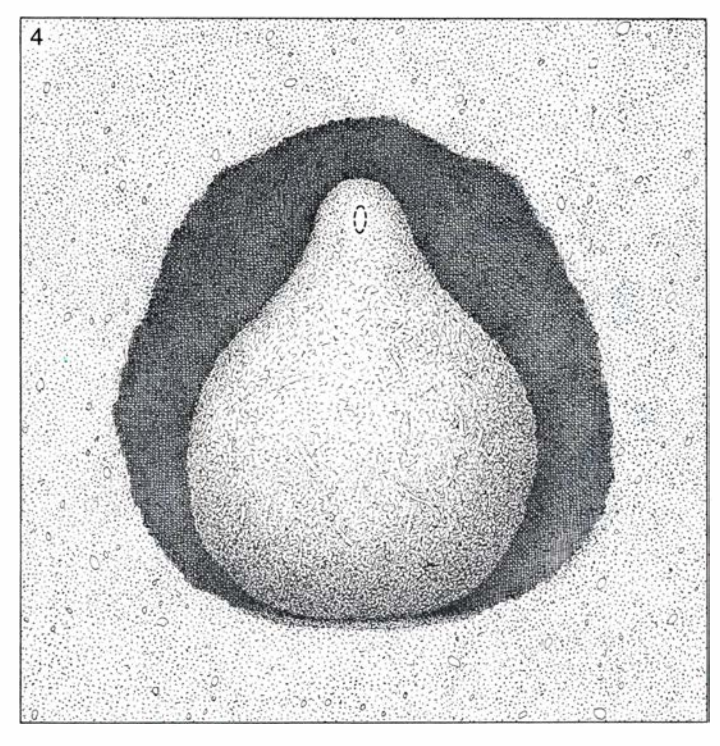
La temperatura de los músculos implicados en el vuelo tiene otras consecuencias; los músculos de vuelo del tórax calientan asimismo otros músculos torácicos que mueven las patas. La actividad de los músculos de vuelo, por tanto, afecta esencialmente toda la actividad de un coleóptero como el escarabajo pelotero: caminar hasta un montón de excrementos después de tomar tierra a cierta distancia del mismo, hacer una pelota, llevársela rodando y defenderla frente a los competidores. Cada una de estas actividades está relacionada con la rapidez de movimiento y la potencia móvil de las patas. Por ello es posible que las temperaturas de vuelo, que evolutivamente se hallan relacionadas con el tamaño del cuerpo, marquen el ritmo de todas las demás actividades, a menos que se pongan en marcha mecanismos fisiológicos para evitarlas.

Los escarabajos grandes poseen un mecanismo fisiológico que les permite liberarse del ambiente térmico. Se trata de la acción de tiritar, que provoca contracciones de los músculos de vuelo. Permite a los escarabajos caldearse para emprender el vuelo, y los mantiene calientes después de haber tomado tierra. Aunque tiritar es esencial para el vuelo, parece ser una opción para otras actividades. Los escarabajos peloteros pueden desplazarse normalmente y hacer otras cosas sobre el suelo con una temperatura corporal próxima al nivel térmico del ambiente. En breve: los escarabajos con la temperatura corporal baja se mueven lentamente.

Descubrimos que la especie diurna *Scarabaeus catenatus*, que en todo caso se mueve con lentitud, tenía una temperatura torácica de 41 grados en vuelo, 28,4 grados mientras moldeaba pelotas a la sombra, 32 cuando las hacía rodar a la sombra y 37 grados cuando hacía rodar las pelotas al sol. La hembra de la especie *Kheper platynotus*, que cabalga sobre la pelota mientras el macho empuja a ésta, muestra siempre una temperatura torácica inferior a la del macho.

Cuando un escarabajo únicamente anda, su temperatura torácica no sube de manera significativa. Por ello llegamos a la conclusión de que los animales tiritan a veces durante determinadas actividades realizadas sobre el suelo a fin





donde enterrarlo (*2*), levantándose sobre sus patas anteriores y empujando la pelota con las posteriores, según es uso entre los escarabajos peloteros. Una vez ha encontrado el punto adecuado, excava (*3*) hasta medio metro en el suelo para enterrar la bola. La hembra pone en ella un solo huevo.

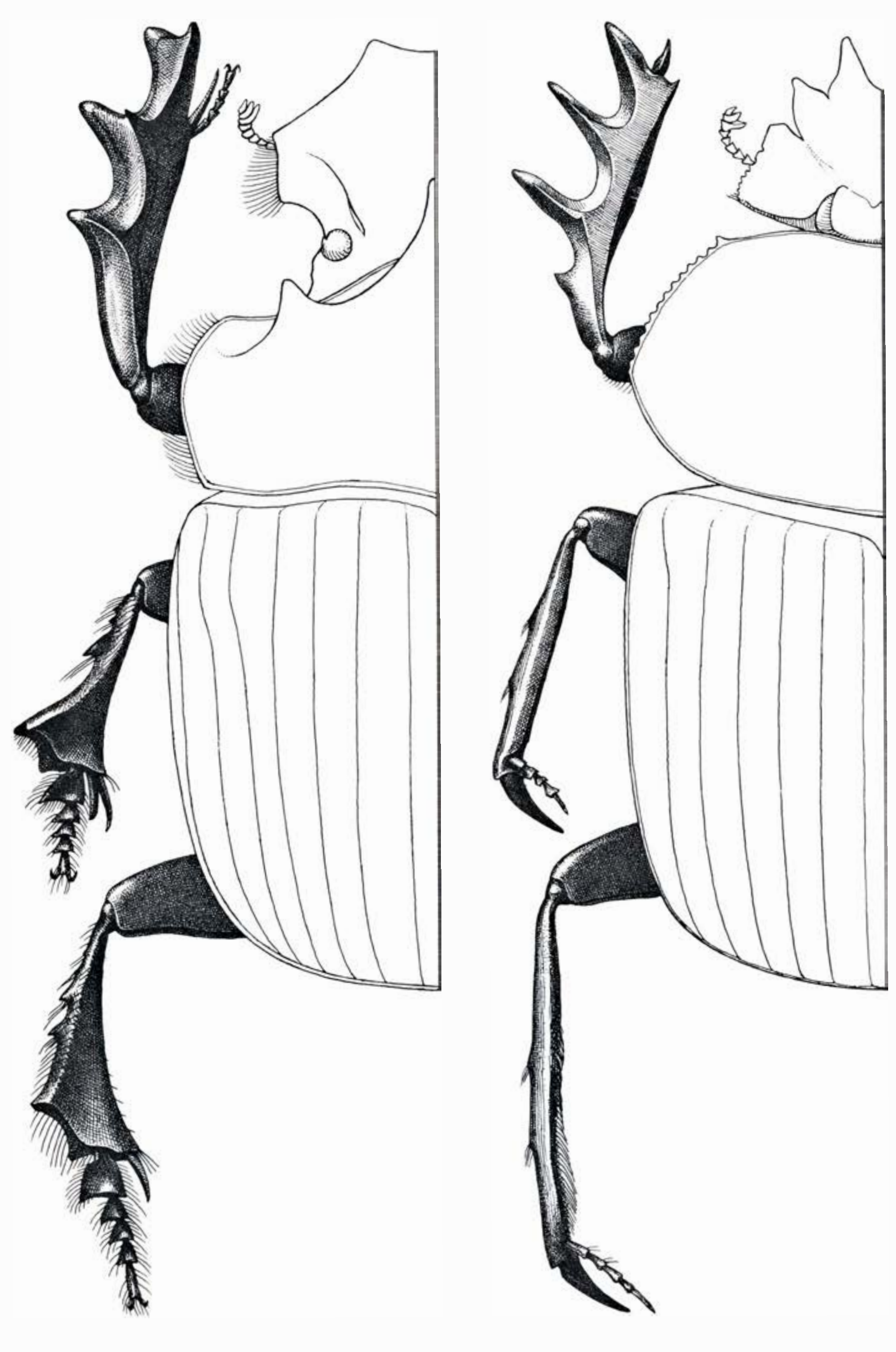
de acelerar su velocidad de trabajo. Sin embargo, debe recordarse que la competencia entre los escarabajos estercoleros activos durante el día es baja, de manera que por lo menos este tipo de presión selectiva, en favor de una aceleración de la tasa de actividad, se ve reducida.

De noche, período en el que los endocópridos y los escarabajos excavadores afluyen en gran número hacia las boñigas, los escarabajos peloteros sólo disponen de unos minutos para procurarse estiércol a partir de los montones frescos o de los que se han acumulado durante el día. Concentramos nuestra atención en una especie de escarabajo pelotero, *Scarabaeus laevistriatus*, que es activa durante la noche. Estos escarabajos se comportaban de forma muy distinta de las especies peloteras que se muestran activas durante el día.

Los escarabajos de la especie *S. laevistriatus*, que son grandes y tiene largas patas, raramente andan de modo pausado, sino que se desplazan a un paso casi frenético. Como otros grandes escarabajos estercoleros del Africa oriental, son voladores poderosos y rápidos, que alcanzan velocidades de hasta 30 kilómetros por hora. Casi con seguridad localizan el estiércol gracias al olfato. Ignoramos si lo buscan volando en sentido perpendicular al viento, desplazándose arriba y abajo de la corriente de aire o bien esperando en el suelo hasta que les llega el olor. En cualquier caso, y a partir de nuestra experiencia, siempre se acercan al estiércol contra el viento. A veces aterrizan directamente sobre las heces, pero más frecuentemente se posan varios metros antes del montón y se escabullen hasta éste corriendo sobre el suelo. Una vez han llegado al estiércol se encaraman rápidamente al mismo, manipulándolo repetidamente con sus patas anteriores y examinándolo con la cabeza, comprobando, aparentemente, si las condiciones de humedad y cohesión lo hacen adecuado para ser moldeado en pelotas.

De aquí en adelante, su comportamiento depende de la magnitud de la actividad de los escarabajos endocópridos y de la presencia o ausencia de otros escarabajos de la especie *S. laevistriatus*. Si el estiércol ha sido extensamente mezclado por los endocópridos, los individuos de *S. laevistriatus* por lo general se van volando al cabo de un minuto o dos, presumiblemente en busca de algo más adecuado. Si no es así, empiezan a construir una pelota. Encuentran una protuberancia sobre la boñiga, la redondean, la cortan, la separan de la masa principal y se la llevan rodando. Algunos excavan en el estiércol, hacen una pelota allí, la empujan hasta la superficie y se la llevan rodando.

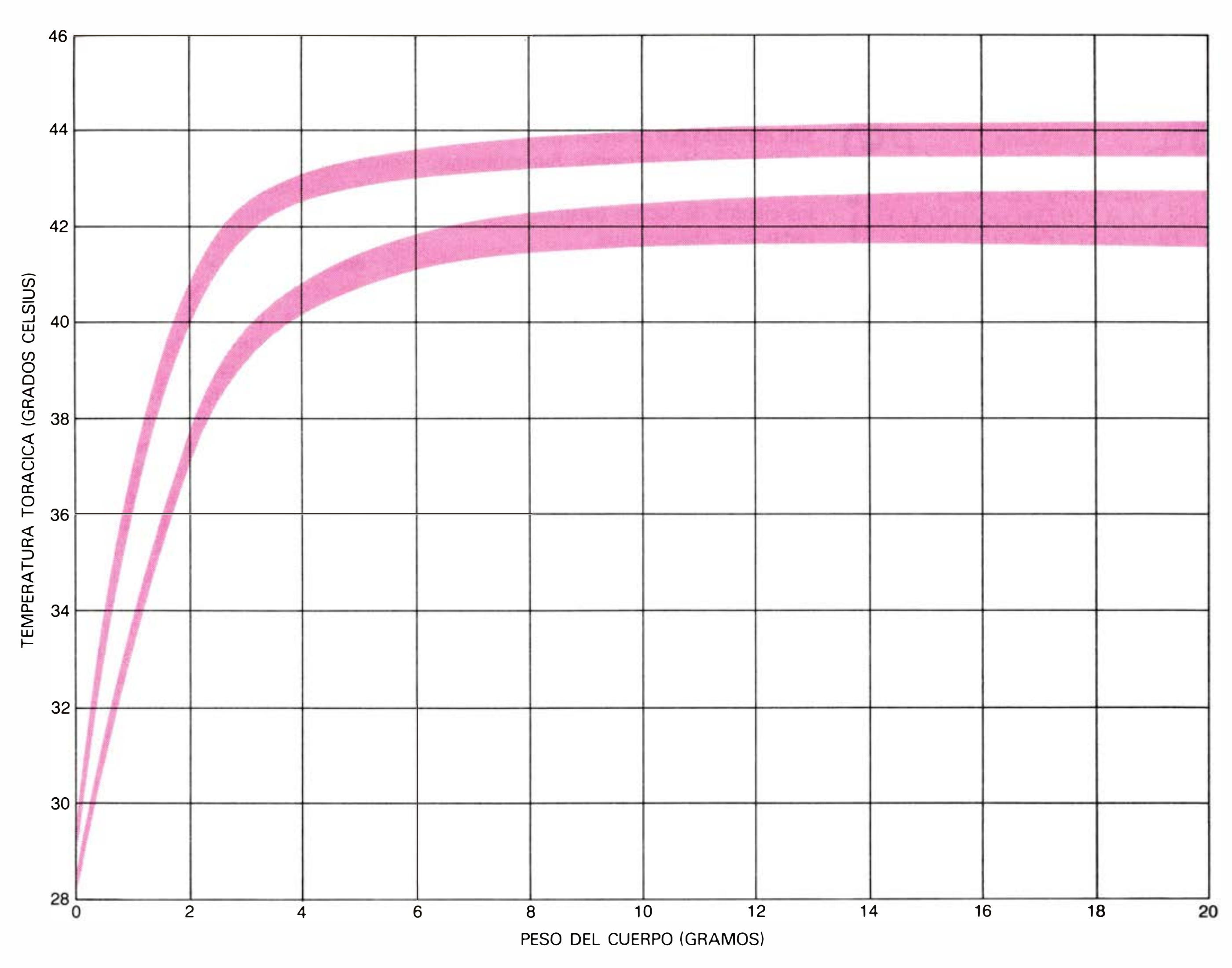
El tiempo que empleaban los escarabajos que observamos en hacer una pelota varió entre 1,1 y 53 minutos. Los períodos más prolongados se observaron únicamente cuando sobre los montones de excrementos colocamos pantallas de alambre para excluir a los endocópridos y a otros competidores. En sólo 15 minutos los endocópridos pueden hacer que los excrementos frescos no resulten ya adecuados para formar pelotas.



LAS ADAPTACIONES MORFOLOGICAS de los escarabajos estercoleros africanos reflejan las distintas maneras que éstos tienen de tratar los excrementos. *Heliocopris dilloni*, que es un escarabajo minador de gran tamaño que se alimenta sólo de excrementos de elefante, presenta las tibias frontales espatuladas, que le ayudan a excavar, y espinas dirigidas hacia atrás en las tibias de las patas posteriores, para aumentar la tracción (*izquierda*). El pelotero *Scarabaeus laevistriatus* (*derecha*) posee patas posteriores largas y delgadas, adecuadas para correr y colgarse del estiércol mientras forma la pelota.

El escarabajo pelotero *S. laevistriatus* utiliza en gran medida sus patas delanteras para añadir fragmentos de estiércol a las pelotas y para moldearlas. La rapidez de los movimientos de golpeteo que se utilizan para ello está directamente relacionada con la temperatura torácica. Los escarabajos diurnos, relativamente fríos, trabajaban con bastante lentitud, pero los movimientos de golpeteo de *S. laevistriatus*, que por lo general estaban calientes, eran extremadamente rápidos.

La relación entre temperatura corpo-



IMPORTANCIA DE LA TEMPERATURA para los escarabajos estercoleros, indicada mediante estas curvas, que representan la temperatura torácica media durante el vuelo (*arriba*) y en el despegue (*abajo*) de varias especies de escarabajos estercoleros africanos, en función de la masa corporal.

ral y velocidad de construcción de la pelota se veía afectada por muchas variables; entre ellas, la consistencia de las heces, el tamaño de la pelota que se estaba construyendo, la densidad de la misma, la presencia o ausencia de otros escarabajos de *S. laevistriatus* que intentaban robar la pelota y la presencia o ausencia de endocópridos que iban deshaciendo la bola conforme se estaba moldeando. Si varios endocópridos invadían una pelota que se hallaba en construcción, los escarabajos de *S. laevistriatus* la abandonaban. (Una pelota de 33 gramos que fue abandonada antes de haber sido terminada contenía más de 50 pequeños endocópridos en su interior.) Pese a estas variables, los escarabajos más calientes eran, por lo general, los primeros en irse con las pelotas terminadas. Constituían una excepción los escarabajos que trabajaban dentro de la pila de excrementos; parecían sacrificar la rapidez de la huida en aras de construir una pelota densa, más cohesionada.

Incluso después de 20 minutos de construcción de pelotas en montones de excrementos que se habían enfriado hasta la temperatura del aire, algunos escarabajos mantenían temperaturas torácicas cercanas a los 40 grados. Basándonos en la velocidad de enfriamiento de los animales muertos, estimamos que si estos escarabajos no hubieran generado calor, en un intervalo de 10 minutos se hubieran enfriado hasta la temperatura del aire o del estiércol, con una variación en más o en menos de dos o tres grados. Por otra parte, algunos escarabajos de la misma especie presentaban temperaturas torácicas inferiores mientras hacían sus pelotas. Ignoramos por qué algunos escarabajos tiritaban y permanecían calientes y otros se enfriaban. Sospechamos que, como pasa en otros animales, la diferencia quizá tenga que ver con la cantidad de energía almacenada de que disponen. Probablemente, los escarabajos que se enfriaban habían agotado ya sus reservas energéticas, y los otros no. El coste energético que implica mantener una temperatura corporal elevada es alto, si bien los beneficios son significativos. Un escarabajo caliente tiene más probabilidades de irse con una pelota de estiércol que un escarabajo frío.

También comprobamos que el éxito de un escarabajo determinado de la especie *S. laevistriatus* en la confección y enterramiento de una pelota de estiércol dependía, en gran parte, de su capacidad de lucha, que resultó estar estrechamente relacionada con la temperatura torácica. La mejor estrategia para obtener rápidamente una pelota era robarla. Un recién llegado a un montón de estiércol suele hacer un intento de este tipo. El escarabajo pelotero maniobra rápidamente para colocarse entre el atacante y la pelota, mientras continúa haciéndola rodar y se aleja. Si, a pesar de todo, el atacante consigue encaramarse a la pelota, tiene lugar una fogosa pelea. Ambos insectos agarran firmemente la pelota con sus patas medias y posteriores e intentan desalojar al antagonista mediante rápidos movimientos hacia adelante y hacia atrás de sus potentes patas anteriores. A veces se sujetan con

sus patas medias y traseras mientras con las delanteras se golpean o empujan. Con frecuencia, el escarabajo perdedor sale arrojado por el aire a unos 10 o más centímetros de distancia. Sin embargo, nunca vimos un contendiente herido en los cientos de luchas naturales y preparadas que observamos.

Para comprobar los efectos de la temperatura sobre este tipo de agresión preparamos peleas con pelotas elaboradas por los escarabajos. Desgraciadamente las pelotas duraban sólo unos cuantos minutos, debido a que los endocópridos penetraban en ellas y las destruían enseguida, y porque el encarnizamiento de la lucha pronto las fragmentaba. Por esas razones construimos pelotas artificiales de arcilla, exprimiendo sobre las mismas excremento fluido de elefante. Los escarabajos recién llegados las aceptaban fácilmente, intentaban hacerlas rodar lejos y las defendían vigorosamente. En nuestros combates preparados observamos tanto conquistas con éxito como fructíferas defensas. Solía vencer el escarabajo de temperatura más elevada. La temperatura corporal elevada resultaba ser más decisiva que el tamaño a la hora de determinar el ganador de un combate.

Al arriesgarse en una lucha, un escarabajo tiene poco que perder, salvo su inversión energética. Las luchas son cortas (duran raramente más de 10 segundos) y producen pocos daños o ninguno. En estas condiciones resulta claramente ventajoso para un escarabajo recién llegado, todavía caliente por el vuelo, arriesgarse a luchar. De lo que se deduce que si un escarabajo ha invertido ya el tiempo y la energía precisos para confeccionar una pelota, debe hacerla rodar lejos y enterrarla tan rápidamente como le sea posible.

Los individuos de la especie *S. laevistriatus* hacían rodar sus bolas de estiércol con presteza. Medimos velocidades altas, de hasta 14 metros por minuto sobre suelo llano, pero sólo en escarabajos con una temperatura torácica de 40 grados o más. La velocidad de rodamiento de las bolas era función directa de la temperatura torácica. Los escarabajos con una temperatura torácica de 42 grados desplazaban las pelotas de estiércol a una velocidad media de 11,4 metros por minuto, mientras que a 32 grados la velocidad media era de 4,8 metros por minuto. En todos los *Scarabaeus laevistriatus* que observamos, la temperatura torácica se elevaba mientras los escarabajos hacían rodar las bolas de estiércol.

Durante esta actividad, la tasa metabólica de los escarabajos endotérmicos es por lo menos tan alta como la de micromamíferos tales como la musaraña, que debe comer una cantidad de alimento al menos igual a su propio peso cada día con el fin de mantener el elevado gasto energético de la endotermia. Cuando el alimento no abunda, muchas aves y mamíferos pequeños abandonan la endotermia, haciéndose temporalmente ectotérmicos y dejando que su temperatura corporal descienda casi hasta la temperatura de su entorno. Los escarabajos de la especie *S. laevistriatus*, nocturnos, se encuentran sobre la cuerda floja energética. La cantidad de energía alimenticia que pueden obtener aumenta con la extensión de su endotermia, pero también aumenta su gasto energético. Pueden maximizar la cantidad neta de energía que obtienen del alimento sólo si la sincronización de su gasto energético es exacta. Sus períodos de endotermia deben coincidir con los momentos en que el rendimiento energético potencial es más elevado.

Puesto que los individuos de *S. laevistriatus* suelen ser activos hasta poco antes del ocaso y durante el mismo, uno se pregunta por qué no se han hecho diurnos. De esta manera se librarían de la intensa competencia a que les someten los endocópridos durante la noche. No conocemos la respuesta. Sólo podemos aventurar que en el pasado evolutivo (y quizás ahora) las aves diurnas y otros depredadores han cobrado un diezmo tan elevado de escarabajos durante las horas de luz que a éstos les es ventajoso ser nocturnos.

Los descubrimientos sobre el papel de la endotermia en los escarabajos estercoleros ayudan a comprender el significado de su evolución en otros animales, incluidos las Aves y los Mamíferos. Los escarabajos estercoleros alcanzan sus tasas de actividad más elevadas sobre el suelo cuando su temperatura corporal se halla al nivel necesariamente generado durante el vuelo; no han desarrollado la capacidad de conseguir tasas de actividad altas a una temperatura corporal baja. En los Coleópteros, como en otros animales, parece que la maquinaria bioquímica se halla adaptada a la más alta temperatura corporal que se alcanza cuando tiene que trabajar a su velocidad máxima, como ocurre con un escarabajo grande en vuelo. De ahí que las máximas tasas de actividad que pueden conseguirse a temperaturas inferiores dependan de los efectos pasivos de la temperatura sobre los procesos metabólicos, que por lo general duplican su velocidad por cada aumento de 10 grados C en la temperatura corporal.

Teoría cuántica y realidad

La doctrina de que el mundo está formado por objetos cuya existencia es independiente de la conciencia humana se halla en conflicto con la mecánica cuántica y con hechos que se han establecido experimentalmente

Bernard d'Espagnat

Cualquier teoría buena en ciencias físicas debe hacer predicciones detalladas. Dado un experimento bien definido, la teoría ha de especificar correctamente el resultado, o al menos debe asignar probabilidades correctas a todos los resultados posibles. Desde este punto de vista, la mecánica cuántica puede considerarse extraordinariamente buena. En su calidad de teoría moderna fundamental de los átomos, de las moléculas, de las partículas elementales, de la radiación electromagnética y del estado sólido suministra métodos para calcular los resultados de la experimentación en todos estos campos.

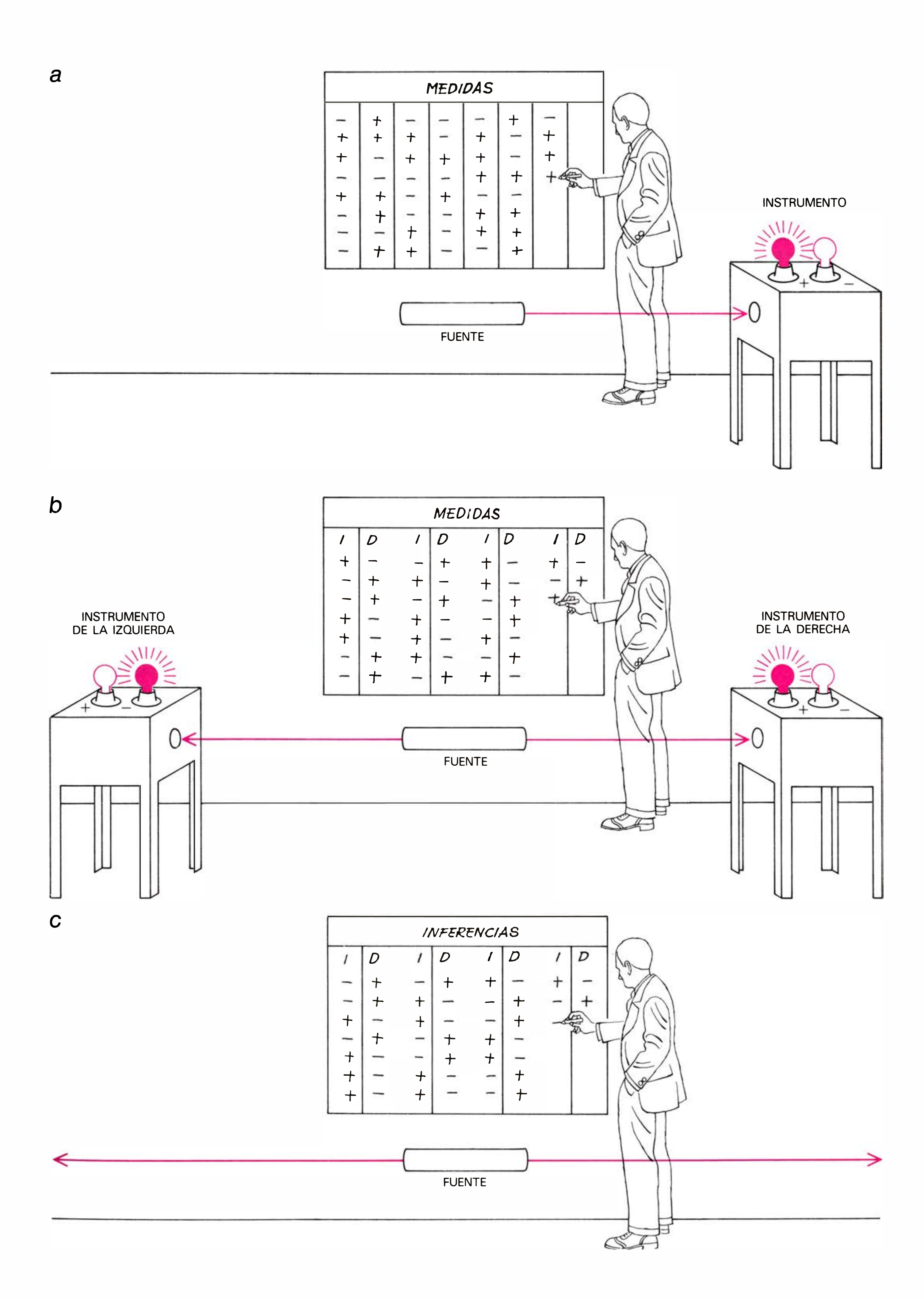
Pero, aparte de una confirmación experimental, podemos pedirle algo más a una teoría. Se espera que no sólo sea capaz de determinar los resultados de un experimento, sino que nos dé también alguna comprensión de los sucesos físicos que presumiblemente sustentan los resultados observados. En otras palabras, la teoría no debe conformarse con dar la posición de una aguja sobre una escala, sino que ha de explicar por qué la aguja toma aquella posición. Cuando se desea información de esta clase en la teoría cuántica surgen algunas dificultades conceptuales. Por ejemplo, en mecánica cuántica una partícula elemental, el electrón, se representa mediante una expresión matemática llamada función de ondas, que frecuentemente describe el electrón como si se hallara esparcido sobre una gran región del espacio.

Esta representación no está en contradicción con la experiencia; por el contrario, la función de ondas da, de forma exacta, la probabilidad de hallar el electrón en un cierto lugar. Sin embargo, cuando el electrón se detecta realmente, nunca está esparcido sino que tiene siempre una posición definida. No está, pues, totalmente claro cuál es la interpretación física que debe asignarse a la función de ondas o qué imagen hemos de formarnos con respecto a qué es un electrón. A causa de estas ambigüedades, muchos físicos encuentran más adecuado considerar la mecánica cuántica como mero conjunto de reglas que permite predecir los resultados de los experimentos. De acuerdo con este punto de vista, la teoría cuántica tratará sólo de los fenómenos observables (la posición observada de la aguja), pero no de cualquier estado físico subyacente (la posición real del electrón).

Resulta ahora que ni siquiera esta renuncia es enteramente satisfactoria. Aceptando incluso que la mecánica cuántica no sea más que un conjunto de reglas, sigue hallándose en conflicto con una imagen del mundo que muchas personas considerarían obvia y natural. Esta imagen del mundo se basa en tres hipótesis, o premisas, que deben aceptarse sin demostración. Una es el realismo, la doctrina que establece que las regularidades apreciadas en los fenómenos observados están causadas por alguna realidad física cuya existencia es independiente del observador. La segunda premisa establece que la inferencia inductiva es una forma válida de razonamiento, que puede aplicarse libremente; por tanto, podemos deducir conclusiones legítimas a partir de observaciones coherentes. La tercera premisa es la llamada separabilidad de Einstein o localidad de Einstein; establece que ninguna clase de influencia puede propagarse más rápidamente que la velocidad de la luz. Las tres premisas, de las que frecuentemente se supone que encierran verdades bien establecidas o incluso verdades totalmente evidentes, forman la base de lo que llamaremos teorías realistas locales de la naturaleza. La argumentación a partir de tales premisas conduce a una predicción explícita de los resultados de una determinada clase de experimentos en física de partículas elementales. También podemos acudir a las reglas de la mecánica cuántica para calcular los resultados de esos experimentos. Ambas predicciones son distintas. Por tanto, o las teorías realistas locales, o la mecánica cuántica, tienen que ser falsas.

Los experimentos en cuestión se propusieron inicialmente como "experimentos imaginarios", esto es, experimentos puramente ideales. Sin embargo, en los últimos años se han llevado a término varias versiones de los mismos con aparatos reales. Aun cuando no todos los resultados son coherentes entre sí, la mayoría están de acuerdo con las predicciones de la mecánica cuántica; y parece que, si coincidencias extraordinarias no han distorsionado los resulta-

CORRELACIONES ENTRE SUCESOS DISTANTES pueden formar las bases de las conclusiones acerca de la estructura del mundo. Supongamos que un físico prepara un dispositivo experimental en el que partículas subatómicas, protones por ejemplo, lanzadas una a una sobre un instrumento, pueden dar sólo dos resultados posibles: más y menos (*a*). Observa que, para algunos protones, el resultado es más y, para otros, es menos; pero no puede decir si el instrumento mide alguna propiedad real de los protones o meramente registra fluctuaciones al azar. El físico puede entonces preparar dos instrumentos idénticos con una fuente que emite dos protones de un modo simultáneo (*b*). Descubre una correlación negativa estricta: siempre que un instrumento lee más, el otro lee menos. A partir de esa correlación, el físico concluye que una propiedad real de los protones es responsable de los resultados y que su valor está determinado antes de que los protones abandonen la fuente. Si la muestra de partículas medidas satisface ciertas pruebas estadísticas, puede inferir que todo par de protones emitidos por la fuente consiste en un protón con la propiedad más y de un protón con la propiedad menos, aun cuando ninguno de los protones se someta al proceso de medición (*c*). Las conclusiones son razonables si se aceptan como válidas tres premisas: que al menos algunas propiedades del mundo tienen una existencia independiente del observador, que la inferencia inductiva puede aplicarse libremente y que una medición hecha con un instrumento no puede influir en el resultado de una medida tomada con el otro instrumento. Una forma más restrictiva de la última premisa prohíbe tales influencias, sólo si las dos mediciones son tan simultáneas que la influencia tendría que propagarse a velocidad superior a la de la luz. Podemos asignar a estas premisas la denominación de realismo, uso libre de la inducción y separabilidad, respectivamente. La versión más restrictiva de la premisa de separabilidad se llama separabilidad o localidad de Einstein. Cualquier teoría que las incorpore es una teoría realista local.


a
MEDIDAS
INSTRUMENTO
+
−
FUENTE
b
MEDIDAS
I
D
INSTRUMENTO DE LA IZQUIERDA
INSTRUMENTO DE LA DERECHA
FUENTE
c
INFERENCIAS
FUENTE

dos, las predicciones de la mecánica cuántica habrán quedado confirmadas. Por consiguiente, las teorías realistas locales son, con gran probabilidad, erróneas. Las tres premisas sobre las que se basan estas teorías son fundamentales para una interpretación del mundo conforme con el sentido común, hasta el punto de que la mayoría de la gente las abandonaría de muy mala gana. Todo parece, sin embargo, que una al menos deberá desecharse, o como mínimo sufrir modificaciones o restricciones en su alcance.

Los experimentos se refieren a las correlaciones entre sucesos distantes y a las causas de las mismas. Sean, por ejemplo, dos partículas que distan entre sí unos metros; supongamos que se descubre que tienen valores idénticos de alguna propiedad, verbigracia la carga eléctrica. Si este resultado se obtiene una vez o unas pocas veces, podemos admitir que se trata de una casualidad, pero si la correlación se detecta de una forma coherente en muchas mediciones, se precisa una explicación más sistemática. No cambiaría las cosas si los valores medidos fueran siempre opuestos en lugar de iguales; la correlación sería entonces negativa, pero su magnitud sería del mismo tamaño, y resultaría igualmente improbable que se debiera al azar.

Cuando se afirma que hay que sobreentender una correlación coherente entre tales sucesos, o que no existe nada misterioso en ello, la explicación ofrecida siempre está relacionada de alguna forma con la causalidad. O bien un suceso origina el otro o ambos sucesos tienen una causa común. Mientras no se da con esa relación, la mente no descansa satisfecha. Más aún, no es posible que lo haga aun cuando se conozcan reglas empíricas para predecir correlaciones futuras. En la antigüedad se observó una correlación entre las mareas y el movimiento de la luna y se formularon reglas para predecir mareas futuras basadas en la experiencia. Pero hasta que Newton introdujo la teoría de la gravitación universal, no era posible sostener que se había llegado a conocer la causa de las mareas.

La necesidad de explicar las correlaciones observadas apremia tanto a los físicos que, a veces, llegan a postular una causa común aunque no exista más pruebas que la misma correlación. Si esta forma de proceder puede o no justificarse constituye el nudo del conflicto entre la mecánica cuántica y las teorías realistas locales. Las correlaciones en cuestión se presentan en observaciones de partículas subatómicas, donde la descripción mecánico-cuántica, con todos sus azares epistemológicos, se hace indispensable. Podemos ilustrar las predicciones de teorías realistas locales a través de la consideración de cómo las correlaciones entre sucesos distantes se explican en un contexto más familiar, donde no es necesario apelar a la mecánica cuántica.

Imaginemos que un psicólogo ha ideado una prueba sencilla, que los sometidos a ella deben superar o fallar, sin que quepa ambigüedad en los resultados. Con los tests delante, el psicólogo descubre que algunos superan la prueba y otros fracasan; pero él no sabe en qué se distinguen ambos grupos, salvo en lo que se refiere a los resultados. En otras palabras, no puede decir que la prueba mida alguna aptitud o capacidad real de los sujetos ni que los resultados sean totalmente fortuitos.

Aunque parece que no hay solución general para este problema, en un caso especial quizá pueda resolverse. Supongamos que la prueba no se presenta a un conjunto de individuos, sino a un conjunto de matrimonios, y que, en las con-

LAS TEORIAS REALISTAS LOCALES y la mecánica cuántica hacen predicciones que están en conflicto en lo relativo a ciertos experimentos en los que se correlacionan sucesos distantes. Las teorías realistas locales predicen que determinada relación, la desigualdad de Bell, se cumplirá, mientras que la mecánica cuántica predice una violación de la misma. Hay prueba experimental muy sólida de que la desigualdad se viola de la forma que predice la mecánica cuántica. Parece pues insostenible la defensa de teorías realistas locales. Una de las hipótesis sobre las que se basan estas teorías debe ser falsa.

testaciones, detectamos una fuerte correlación. El procedimiento puede consistir en separar, antes del test, los maridos de las mujeres y, luego, someterlos a la prueba por separado. Al analizar los resultados, volvemos a observar que una parte de la población ha respondido bien y otra parte ha fallado, con la peculiaridad de que en aquellos casos en que el marido "aprobó" lo hizo también su mujer y cuando fracasó el marido tampoco pasó su mujer.

Si la correlación persiste después de haber realizado la prueba con muchos matrimonios, el psicólogo puede concluir con gran seguridad que la respuesta de cada individuo no se debía al azar en el momento del test. Por el contrario, la prueba ha de revelar alguna propiedad o habilidad real de los individuos. La propiedad deben poseerla los individuos antes del ejercicio y antes de haberlos separado. El azar pudo incidir de algún modo en el desarrollo de la propiedad, ya que no todas las parejas la poseían, pero esa incidencia debió actuar en algún instante antes de la separación por sexos. En ese período previo, en que los maridos y mujeres estaban juntos, fue cuando pudieron adquirir alguna característica que les habría de permitir contestar coherentemente de la misma forma. Así pues, la correlación queda explicada atribuyéndola a una causa común anterior a la prueba.

Otra explicación que debe excluirse antes de llegar a esta conclusión es que los maridos y las mujeres se hubieran comunicado entre sí durante la realización del ejercicio. Si hubiera habido algún medio de comunicación, no habría necesidad de que existiera una habilidad común antes de la prueba. Cualquiera de los esposos que hubiera realizado la prueba en primer lugar hubiera podido escoger la respuesta al azar y haber mandado instrucciones al otro, creando así la correlación observada. Al contestar un test psicológico no es difícil evitar subterfugios de esta clase. En el caso extremo, las pruebas podrían realizarse en estricta simultaneidad y los maridos y mujeres podrían instalarse en lugares tan alejados que ninguna señal que se moviera a velocidades inferiores a la de la luz pudiera llegar a tiempo para que tuviera alguna utilidad.

Una vez aclarado que la prueba mide alguna propiedad real, el psicólogo puede dar un paso adelante y sacar una inferencia inductiva. Si las parejas que ya han sufrido la prueba constituyen una muestra no sesgada de una población de parejas, y si la muestra satisface ciertas condiciones estadísticas, el psicólogo puede inferir que cualquier pareja de la misma población estará formada por un marido y una mujer que o bien ambos poseen o bien ambos no poseen la propiedad medida por la prueba. Por el mismo principio puede concluir que, en cualquier muestra grande y no desviada de parejas que aún no han realizado la prueba, habrá matrimonios que tendrán la propiedad y otros que carecerán de ella. La seguridad de estas afirmaciones se va acercando al estado de certeza a medida que aumenta el tamaño de la muestra. Por consiguiente, se infiere que tanto la correlación dentro de las parejas como la existencia de diferencias entre las parejas persisten incluso en la fracción de población que no se ha sometido a la prueba.

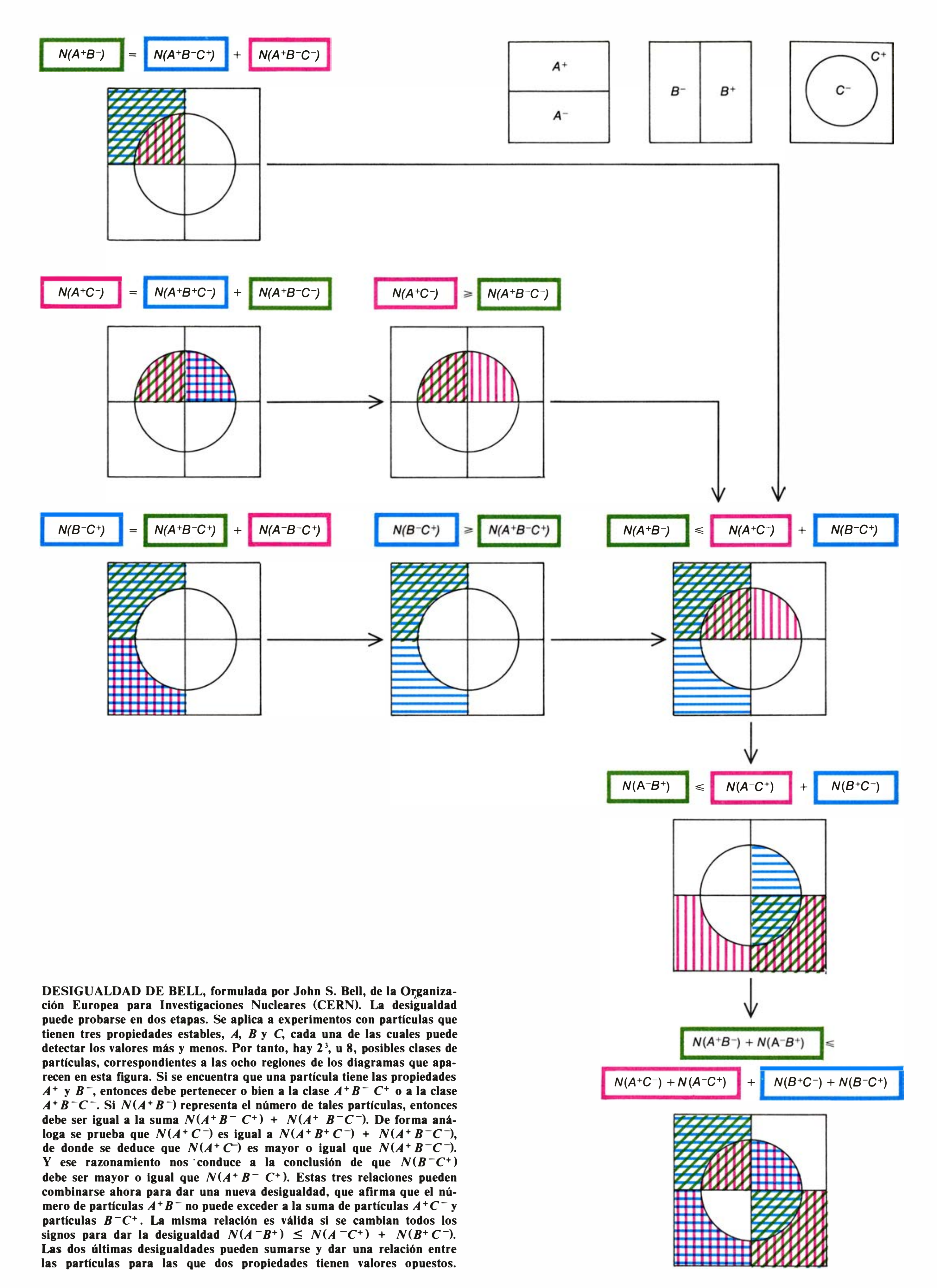
Estas conclusiones están basadas en las tres premisas que constituyen el fundamento de las teorías realistas locales. El realismo es una hipótesis necesaria si creemos que algunas pruebas miden propiedades estables que existen independientemente del experimentador. Fue necesario suponer la validez de la inferencia inductiva para extrapolar los datos observados a la parte de la población que no había realizado el test. La separabilidad se incorporó en la hipótesis de que los maridos y las mujeres sometidos al ejercicio no podían comunicarse entre sí. Si las pruebas se desarrollaban en estricta simultaneidad, de forma que cualquier señal que pasara entre los maridos y las mujeres debería propagarse a velocidad mayor que la de la luz, la hipótesis resultaba equivalente a la separabilidad de Einstein.

A primera vista, las conclusiones extraídas de esta experiencia hipotética en el ámbito de la psicología parecen deducirse de forma totalmente natural de los datos. Un epistemólogo podría objetar, sin embargo, que las conclusiones son inciertas. En particular, un epistemólogo que conociera los fundamentos de la mecánica cuántica podría argumentar que ninguna necesidad lógica nos obliga a aceptar las tres premisas del razonamiento del psicólogo; por consiguiente, tampoco sería necesario concluir que existía una correlación entre maridos y mujeres antes de someterse al test, ni que había diferencias entre las parejas antes de realizar la prueba. Al psicólogo no le parecerán serias, a buen seguro, esas objeciones y las considerará expresión de una duda infundada o una creencia, muy poco científica, en una paradoja. En la bibliografía relativa a la mecánica cuántica hallamos numerosos argumentos similares, o equivalentes en su forma, a éste, todos ellos encaminados a probar que las correlaciones o diferencias no tenían por qué existir antes de ser medidas.

Característica singular de la mecánica cuántica es que sus predicciones suelen ofrecer sólo la probabilidad de un resultado, no la afirmación determinista de que el suceso tendrá o no lugar. La función de onda empleada para describir el movimiento de una partícula elemental se interpreta con frecuencia de forma probabilística: la probabilidad de encontrar la partícula en un cierto punto es proporcional al cuadrado de la función de ondas en ese punto. Como expuse antes, una función de onda puede hallarse a veces esparcida sobre una gran región; ello implica que la probabilidad pueda estar ampliamente distribuida. Por supuesto, cuando se realiza una medición en un punto determinado, la partícula se detecta o no se detecta; se habla entonces de un colapso de la función de ondas. Supongamos que se detecta la partícula. Desde el punto de vista epistemológico la cuestión que interesa dilucidar entonces será: ¿Ocupaba la partícula esa posición antes incluso de haberse realizado la medición?

Las conclusiones del psicólogo, si pudieran trasladarse a este contexto, implicarían que la posición de la partícula estaba bien definida desde el principio, de la misma forma que la habilidad descubierta en algunos miembros de la muestra preexistía a la realización del test. De acuerdo con este razonamiento, la posición de la partícula nunca estuvo indeterminada; ocurría sencillamente que era desconocida para el observador.

Gran parte de la plana mayor de los físicos expertos en mecánica cuántica estarían en desacuerdo. Pero no todos; Einstein se mostró, a lo largo de su vida, reticente ante la naturaleza probabilística de las interpretaciones que solían darse en mecánica cuántica. La mayoría de sus críticas incisivas a esas interpretaciones se fundaban en un razonamiento que se asemeja en cierto sentido al que yo he atribuido al psicólogo. En 1935 Einstein publicó un trabajo con dos jóvenes colegas, Boris Podolsky y Nathan Rosen, en el que formuló explícitamente sus objeciones. No afirmaba que la teoría cuántica fuera falsa; por el contrario, suponía que algunas, por lo menos, de sus predicciones debían ser correctas. Lisamente proponía que la descripción mecánico-cuántica de la naturaleza resultaba incompleta o aproximada. El movimiento de una partícula debe describirse en términos de probabilidades, decía, por la única razón de que algunos de los parámetros que determinan el movimiento todavía no han sido

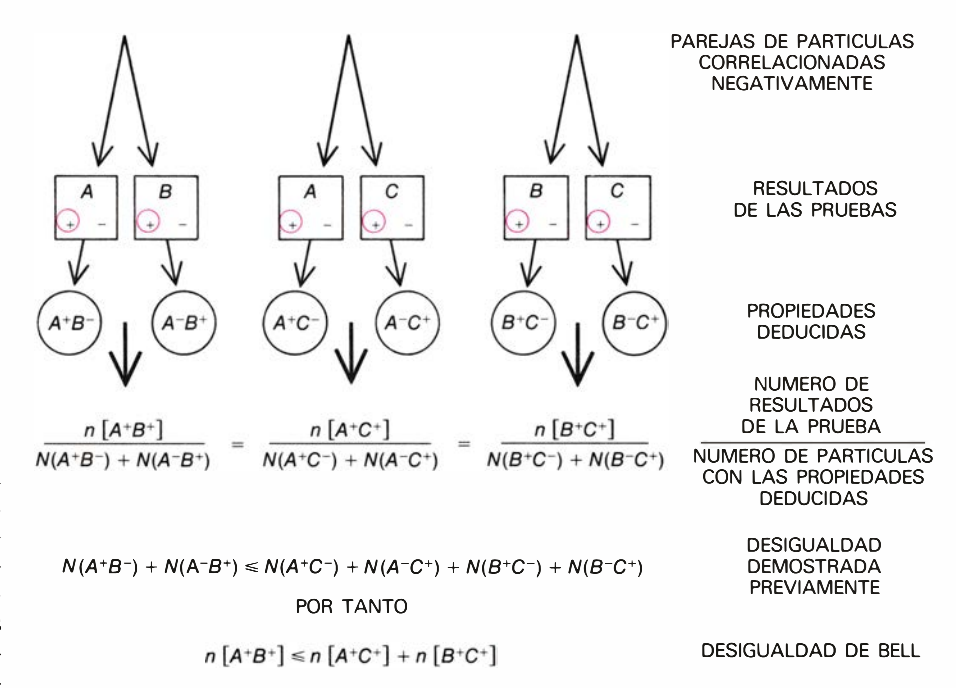


DESIGUALDAD DE BELL, formulada por John S. Bell, de la Organización Europea para Investigaciones Nucleares (CERN). La desigualdad puede probarse en dos etapas. Se aplica a experimentos con partículas que tienen tres propiedades estables, *A*, *B* y *C*, cada una de las cuales puede detectar los valores más y menos. Por tanto, hay 2^3, u 8, posibles clases de partículas, correspondientes a las ocho regiones de los diagramas que aparecen en esta figura. Si se encuentra que una partícula tiene las propiedades A^+ y B^-, entonces debe pertenecer o bien a la clase $A^+B^-C^+$ o a la clase $A^+B^-C^-$. Si $N(A^+B^-)$ representa el número de tales partículas, entonces debe ser igual a la suma $N(A^+B^-C^+) + N(A^+B^-C^-)$. De forma análoga se prueba que $N(A^+C^-)$ es igual a $N(A^+B^+C^-) + N(A^+B^-C^-)$, de donde se deduce que $N(A^+C^-)$ es mayor o igual que $N(A^+B^-C^-)$. Y ese razonamiento nos conduce a la conclusión de que $N(B^-C^+)$ debe ser mayor o igual que $N(A^+B^-C^+)$. Estas tres relaciones pueden combinarse ahora para dar una nueva desigualdad, que afirma que el número de partículas A^+B^- no puede exceder a la suma de partículas A^+C^- y partículas B^-C^+. La misma relación es válida si se cambian todos los signos para dar la desigualdad $N(A^-B^+) \leq N(A^-C^+) + N(B^+C^-)$. Las dos últimas desigualdades pueden sumarse y dar una relación entre las partículas para las que dos propiedades tienen valores opuestos.

especificados. En cuanto los valores de estas hipotéticas "variables ocultas" lleguen a conocerse, se podrá definir una trayectoria totalmente determinista.

Contra la propuesta de Einstein se han formulado numerosos contraargumentos. Por ahora mencionaré sólo uno de ellos, que se funda en el criterio de utilidad. El argumento establece que es irrelevante que existan o no variables ocultas, que se den diferencias entre los matrimonios en ausencia de test. Aun cuando existan, no deberían incorporarse en ninguna teoría ideada para explicar las observaciones; podemos decir, por tanto, que carecen de existencia científica. La exclusión de las variables ocultas queda justificada por la conjunción de tres hechos. Primero, el formalismo matemático de la teoría se simplifica si prescindimos de las variables ocultas. Segundo, este formalismo simplificado predice resultados que confirma la experimentación. Tercero, la adición de variables ocultas a la teoría no originaría nuevas predicciones que pudieran verificarse. Por tanto, la afirmación de que existen variables ocultas trasciende el alcance de los experimentos y no es una proposición de la física, sino de la metafísica.

Esta defensa de la interpretación clásica de la mecánica cuántica elimina cualquier variable oculta por superflua y, en última instancia, quizá, por sin sentido. Desarrollos teóricos recientes han demostrado que la situación actual es muy distinta. La hipótesis de que existen variables ocultas conduce de hecho a predicciones experimentales que difieren de las previsiones a que llegaba la mecánica cuántica. Teorías con variables ocultas, y teorías realistas locales, en general, ponen límites a la extensión hasta donde ciertos sucesos distantes pueden hallarse correlacionados; la mecánica cuántica, por el contrario, predice que, en algunas circunstancias, el límite puede superarse. Por tanto, debería ser posible, al menos en principio, idear una prueba experimental que discriminara entre las dos teorías.



EN LA SEGUNDA ETAPA DE LA PRUEBA se extrapola desde el caso de partículas únicas para las que se conocen dos propiedades hasta el caso de pares de partículas, en cada una de las cuales se mide una sola propiedad. Estos pares se crean de suerte que siempre existe una correlación negativa estricta para cualquier propiedad considerada por separado, esto es, si una partícula en el par tiene la propiedad A^+, la otra debe tener la propiedad A^-. Debido a esta correlación, si una partícula de un par tiene la propiedad A^+, y se halla que la otra posee la propiedad B^+, es posible deducir ambas propiedades de las dos partículas. Una prueba doblemente positiva se puede originar sólo si una de las partículas tiene las dos propiedades A^+B^- y la otra las dos propiedades A^-B^+. Por tanto, el número de tales pruebas con resultados positivos dobles, que se designará por $n[A^+B^+]$, debe ser proporcional al número total de partículas con las propiedades A^+B^- y A^-B^+. Pueden derivarse proporcionalidades parecidas para el número de resultados positivos dobles observados cuando se miden en pares de partículas las propiedades A y C y las propiedades B y C; estas son las cantidades $n[A^+C^+]$ y $n[B^+C^+]$. La constante de proporcionalidad depende sólo del número de pares sometido a cada conjunto de pruebas y del número total de casos; por tanto, la constante será la misma en los tres casos. Síguese que los tres cocientes del número de resultados de la prueba que son doblemente positivos dividido por el número de partículas individuales que pueden dar origen a estos resultados también serán iguales. Ya se ha demostrado una relación entre los números de partículas individuales con las propiedades indicadas. Se trata de la desigualdad probada en la figura anterior. Si aquella desigualdad es cierta, debe existir una desigualdad análoga entre los números de resultados de pruebas doblemente positivas. Y esa es la desigualdad de Bell. La prueba es válida si las tres premisas de las teorías realistas locales se suponen válidas.

Supongamos que un físico ha ideado una demostración que puede efectuarse con partículas subatómicas, protones por ejemplo. Tras muchos intentos, descubre que unos protones pasan la prueba y otros fallan; pero él no sabe si está midiendo alguna propiedad real de los protones o meramente observando fluctuaciones al azar de su aparato. Y por ende, intenta aplicar la prueba a pares de protones, no a protones individuales. Los protones que constituyen un par están inicialmente muy próximos, acercados por un procedimiento bien definido que es el mismo para todos los pares. Se permite luego que los protones se separen; cuando se han alejado cierta distancia macroscópica, se les somete a prueba, simultáneamente para algunos pares y con un intervalo de tiempo entre pruebas para los pares restantes. El físico descubre una estricta correlación negativa: cuando, en un par, un protón pasa la prueba, el otro falla invariablemente.

La situación del físico tiene analogías obvias con las del psicólogo que realiza el test con parejas; y puede aplicarse el mismo razonamiento a los resultados del experimento físico. Si se aceptan como premisas el realismo, el uso libre de la inducción y la separabilidad de Einstein, el físico se sentirá justificado para concluir que la prueba mide alguna propiedad real de los protones. Para que la correlación pueda explicarse, la propiedad debe preexistir a la separación de los protones de cada par y ha de tener algún valor definido para ellos desde el momento en que exista hasta que se lleve a cabo el experimento. Todavía más. Si se preparan nuevos pares de protones por el mismo procedimiento, el físico sabrá que, en cada caso, un protón tendrá la propiedad y el otro no, aun cuando no se someta a prueba ninguno de los protones.

¿Hay alguna prueba real que pueda acometerse con partículas subatómicas y que produzca resultados análogos? Existe. Se trata de la medición de cualquier componente, definida a lo largo de algún eje arbitrario, del spin de la partícula. El spin atribuido a una partícula subatómica es análogo sólo en algunos aspectos al momento angular de rotación de un cuerpo macroscópico, el de la tierra, por ejemplo. Sin embargo, para esta discusión no necesitamos introducir los detalles de cómo se trata el spin en mecánica cuántica. Basta decir que el spin de una partícula se representa mediante un vector, o flecha, que podemos

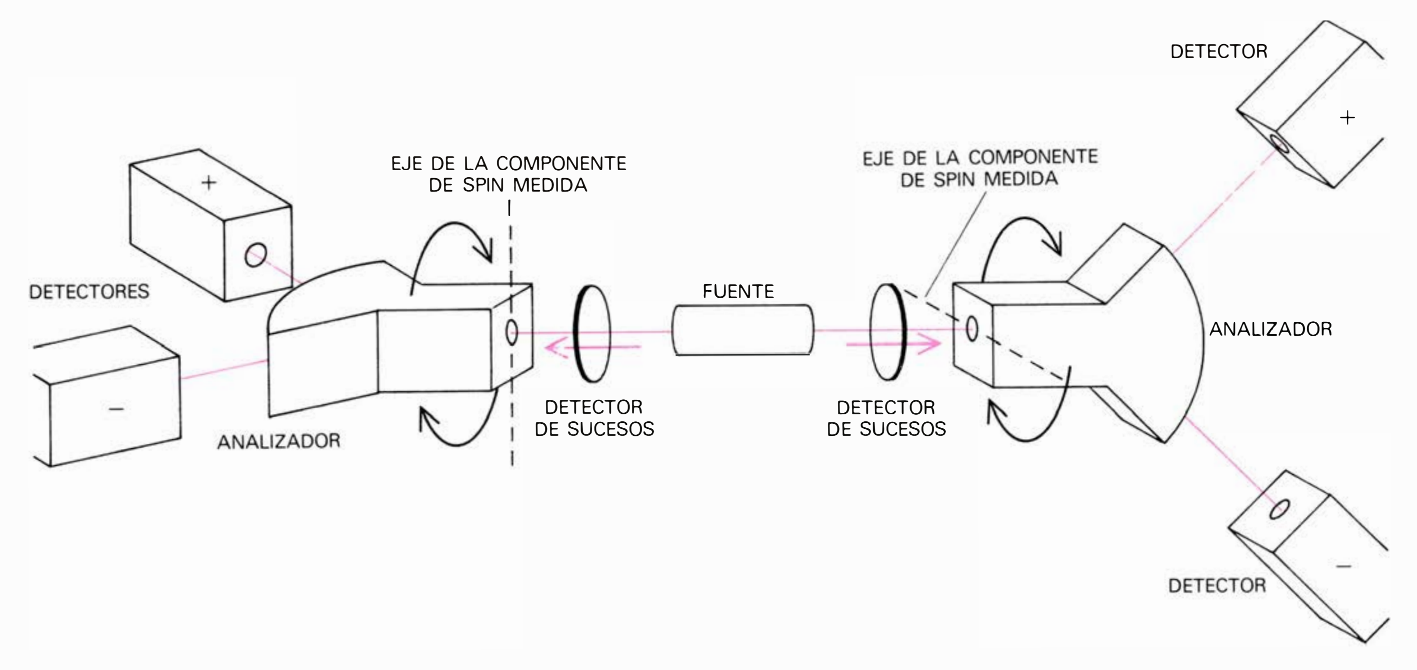
imaginar ligado a la partícula. Una proyección de este vector sobre cualquier eje en el espacio tridimensional es la componente del spin a lo largo de ese eje. Una propiedad bien establecida, aunque no menos sorprendente, de los protones (y de muchas otras partículas) es que, cualquiera que sea el eje elegido para medir la componente del spin, los resultados pueden tomar únicamente dos valores, que llamaré más y menos. (La medición de la componente del momento angular de rotación de la tierra daría distintos resultados, según la dirección de la componente; y tendría cualquier valor, desde cero hasta el momento angular total de la tierra.)

Se observa una correlación estrictamente negativa entre las componentes del spin cuando se juntan dos protones en la configuración mecánico-cuántica llamada estado singlete. En otras palabras, si dejamos separar dos protones en estado singlete y se mide luego la misma componente del spin en ambas partículas, será siempre más para un protón y menos para el otro. No hay forma conocida de predecir qué partícula tendrá la componente más y cuál poseerá la componente menos; sin que ello sea obstáculo para que la correlación negativa esté bien establecida. Ni cambia la situación sea cual sea la componente del spin que el físico decida medir, con tal que en ambas partículas se mida la misma componente. Tampoco importa cuánto se han alejado los dos protones antes de medir, mientras no se presenten influencias perturbadoras, tales como otras partículas o radiación, a lo largo de sus caminos.

En lo concerniente a esta sencilla medición no hay razón de conflicto entre las predicciones de la mecánica cuántica y las de las teorías realistas locales. Pero pueden aparecer las discrepancias si el experimento se complica algo más.

El vector que representa el spin de una partícula se define mediante las componentes a lo largo de tres ejes en el espacio, que no forman necesariamente ángulos rectos entre sí. Para un vector asociado con un objeto macroscópico de la vida normal, podría darse por supuesto, y con toda razón, que las tres componentes tienen valores definidos en cualquier instante. Que desconozcamos quizás el valor de una componente, no es lo mismo que pueda estar indefinido. Sin embargo, cuando se aplica esta hipótesis al vector de spin de una partícula resulta altamente sospechosa; se desecha, de hecho, en la interpretación normal de la mecánica cuántica, como un ejemplo de teoría con variables ocultas. El problema es que no se puede idear, ni siquiera en el terreno de los principios, ningún experimento que aportara información sobre los valores simultáneos de las tres componentes. Un aparato está capacitado para medir únicamente una componente del spin. Y, al hacerlo, altera de ordinario los valores de las otras componentes. Por tanto, para conocer los valores de las tres componentes deberían hacerse tres mediciones sucesivas. Cuando la partícula saliera del tercer aparato no tendría ya las mismas componentes del spin que cuando entró en el primer aparato.

Aunque los instrumentos sólo pueden medir una componente del spin cada vez, sí podemos construir un aparato que mida la componente del spin a lo largo de uno de los tres ejes elegidos arbitrariamente. Designaré estos ejes por A, B y C e indicaré los resultados de los experimentos como sigue: Si se encuentra que la componente del spin a lo largo del eje A es más, se indicará por A^+. Si la componente a lo largo del eje B es menos, escribiremos B^-, y así sucesivamente. El físico puede preparar ya una gran muestra de protones en el estado singlete. Observará que si mide la componente A para ambos protones de un par, unos tendrán A^+ y otros tendrán A^-, pero siempre que un miembro del par sea A^+, el otro miembro será siempre A^-. Si decide medir, en vez de la



EXPERIMENTO IMAGINARIO para comprobar la desigualdad de Bell, midiendo las componentes del spin de los protones u otras partículas elementales. Una componente del spin es la proyección según el eje del momento angular intrínseco del protón. Cada componente puede tener únicamente dos valores, que llamaremos más y menos. El experimento, que supone que podemos disponer de instrumentos perfectos, tendría una fuente donde se reunirían pares de protones en una configuración mecánico-cuántica conocida por estado singlete. A continuación, los pares se romperían; los protones saldrían en direcciones opuestas. "Los detectores" suministrarían una señal cada vez que se emitiera un par adecuado de protones. Cada protón entraría, entonces, en un analizador; éste lo desviaría hacia uno de los dos detectores según el valor de la componente de su spin a lo largo del eje definido por el analizador. Si los analizadores se colocaran para medir las componentes del spin a lo largo del mismo eje, se observaría una correlación negativa estricta. Si se girara uno de los analizadores, de forma que se midieran componentes distintas, las teorías realistas locales predicen que las correlaciones observadas no serían mayores que las permitidas por la llamada desigualdad de Bell, con independencia de cuál fuera el ángulo entre los analizadores. La mecánica cuántica predice una violación de la desigualdad de Bell para algunos ángulos. (Ilustraciones de Jerome Kuhl.)

componente A, la componente B, hallará la misma correlación negativa: cuando un protón sea B^+, su compañero del singlete será B^-. De modo parecido, un protón C^+ está invariablemente acompañado por un C^-. Estos resultados son válidos independientemente de cómo estén orientados los ejes A, B y C.

Importa destacar que, en estos experimentos, no se somete ningún protón a una medición de más de una componente de su spin. Pero, si el físico acepta las tres premisas de las teorías realistas locales, puede deducir, a partir de sus resultados, conclusiones acerca de los valores de las tres componentes, siguiendo un razonamiento muy parecido al del psicólogo de la hipótesis del principio. Considerando un nuevo grupo de pares de protones en el estado singlete, en el que no se han realizado todavía mediciones del spin (y en el que quizá nunca se tomará tales mediciones), puede inferir que, en cada par, un protón tiene la propiedad A^+ y el otro la propiedad A^-. De forma análoga, puede concluir que, en cada par, un protón goza de la propiedad B^+ y el otro de la B^- y que uno muestra la propiedad C^+ y el otro la C^-.

Estas conclusiones exigen una sutil, e importante, extensión del significado asignado a una notación del tipo A^+. Antes A^+ sólo era un posible resultado de una medición de la partícula; pero ahora se ha convertido, merced a este razonamiento, en un atributo de la partícula misma. Para ser explícitos, si un protón no sujeto a medición detenta la propiedad de que una medición a lo largo del eje A dará el resultado definido por A^+, entonces se dice que el protón tiene la propiedad A^+. En otras palabras, el físico se ha visto forzado a concluir que los protones de cada par poseen componentes del spin bien definidas en cada instante. Podemos desconocer las componentes, ya que el físico no puede decir qué protón del par tiene la propiedad A^+ y cuál la propiedad A^- en tanto no haya realizado la correspondiente medición a lo largo del eje A, pero puede argumentar, a partir de las premisas de teorías realistas locales, que los valores están bien definidos incluso en ausencia de cualquier tipo de medición. Este punto de vista es contrario a la interpretación convencional de la mecánica cuántica, pero ningún hecho de los que se han ido introduciendo la ha contradicho todavía.

Se espera una correlación negativa estricta para protones en el estado singlete en el caso exclusivo en que se mide la misma componente del spin para ambos protones. ¿Qué sucede cuando los instrumentos se disponen para que midan distintas componentes del spin? Para precisar, consideremos el siguiente experimento. Se juntan pares de protones en un estado singlete por el mismo método empleado en los experimentos anteriores; se permite su separación bajo las mismas condiciones exactamente. En cada protón, medimos una componente del spin, A, B o C, pero se determina de forma totalmente al azar qué componente vamos a medir en cada caso. A veces se medirá la misma componente del spin en ambos protones; pero tales resultados se eliminan, por no aportar nueva información. Los pares restantes constarán, entonces, de un protón en el que la medición se ha tomado a lo largo del eje A y otro a lo largo del eje B, o de una medición a lo largo del eje A y el otro a lo largo del eje C, o uno a lo largo del eje B y el otro a lo largo del C. Para simplificar, denominaré los pares de cada una de estas tres muestras por AB, AC y BC. Un par que al ser medido dé el resultado A^+ para un protón y el B^+ para el otro puede representarse por A^+B^+. El número de pares observados de este tipo puede representarse por $n[A^+B^+]$. ¿Cabe esperar una relación entre estas cantidades?

En 1964 John S. Bell, de la Organización Europea de Investigaciones Nucleares (CERN), descubrió esta relación. Para cualquier gran muestra de pares de protones en estado singlete Bell demostró que las hipótesis de las teorías realistas locales imponían un límite en la correlación que podía esperarse cuando se medían distintas componentes del spin. El límite se expresa en forma de una desigualdad, que ahora se llama desigualdad de Bell. Dadas las condiciones experimentales mencionadas antes establece que el número de pares A^+B^+ no puede exceder a la suma del número de pares A^+C^+ y el número de pares B^+C^+. La desigualdad puede simbolizarse en la expresión siguiente:

$$n[A^+B^+] \leq n[A^+C^+] + n[B^+C^+].$$

Podríamos construir muchas desigualdades similares transponiendo los símbolos o cambiando los signos. Como las direcciones a lo largo de las cuales se definen las componentes del spin se eligieron de un modo arbitrario, todas estas formulaciones son intercambiables. Me detendré sólo en la anterior.

La desigualdad de Bell puede ser demostrada dentro del contexto de las teorías realistas locales, mediante un simple razonamiento de la teoría matemática de conjuntos. Es útil comenzar con una hipótesis contraria a los hechos: que existe alguna forma de medir independientemente dos componentes del spin de una partícula dada. Supongamos que este aparato inexistente ha revelado que un determinado protón tiene componentes de spin A^+ y B^-. La tercera componente, C, no se ha medido, pero sólo puede tener dos valores: más o menos. Por tanto, el protón considerado debe ser un miembro de uno de los dos conjuntos de protones, o bien del conjunto con componentes de spin $A^+B^-C^+$ o bien del conjunto con componentes $A^+B^-C^-$. No hay otras posibilidades.

Si se detectan muchos protones con componentes de spin A^+B^-, se puede escribir una ecuación relativa:

$$N(A^+B^-) = N(A^+B^-C^+) + N(A^+B^-C^-).$$

Al objeto de evitar confusiones se ha usado el formalismo $N(A^+B^-)$ para representar el número de protones individuales con las dos componentes del spin A^+ y B^-; el símbolo $n[A^+ \; B^-]$ da el número de pares de protones en los que una partícula tiene la componente A^+ y, la otra, la componente B^-. La ecuación establece el hecho evidente de que, cuando un conjunto de partículas se divide en dos subconjuntos, el número total de partículas del conjunto original debe ser igual a la suma del número de partículas de los subconjuntos.

Los protones que aparecen dotados de componentes de spin A^+C^- pueden analizarse de forma análoga. Todo protón de este tipo debe ser miembro del conjunto $A^+B^+C^-$ o del conjunto $A^+B^-C^-$; el número total $N(A^+C^-)$ tendrá que equivaler a la suma $N(A^+B^+C^-)$ + $N(A^+B^-C^-)$. Podemos adelantar un paso más. Si el número de protones $N(A^+C^-)$ es igual a $N(A^+B^+C^-)$ + $N(A^+B^-C^-)$, entonces debe ser mayor o igual que $N(A^+B^-C^-)$. [Los dos conjuntos serán iguales si las componentes B del spin de todas las partículas sean menos, de forma que el subconjunto $(A^+B^+C^-)$ esté vacío; en caso contrario, $N(A^+C^-)$ resultará mayor. En otras palabras, la parte no puede ser mayor que el todo.) Podemos recurrir de nuevo al mismo razonamiento para probar que el número de protones con componentes de spin B^-C^+ debe ser igual a la suma $N(A^+B^-C^+) + N(A^-B^-C^+)$ y, por tanto, $N(B^-C^+)$ debe ser mayor o igual que $N(A^+B^-C^+)$.

Consideremos de nuevo la primera ecuación derivada antes

$$N(A^+B^-) = N(A^+B^-C^+) + N(A^+B^-C^-).$$

Acabamos de probar que $N(B^-C^+)$ es mayor o igual que $N(A^+B^-C^+)$, que es el primer término del miembro de la derecha de esta ecuación. Se ha demostrado también que $N(A^+C^-)$ es mayor o igual que $N(A^+B^-C^-)$, que es el segundo miembro del término de la derecha de la ecuación. Cabe, pues, hacer las sustituciones apropiadas en la ecuación, cambiando el signo "igual" por otro que signifique "menor o igual que".

El resultado da la desigualdad

$$N(A^+B^-) \leq N(A^+C^-) + N(B^-C^+).$$

Aunque esta desigualdad se ha derivado aquí formalmente, no puede comprobarse de una manera directa por vía experimental, porque no existe aparato alguno capaz de medir independientemente las dos componentes del spin de un único protón. Pero los experimentos a que nos estamos refiriendo no se realizan con protones individuales, sino con pares correlacionados de los mismos; no es, pues, necesario realizar tales mediciones imposibles. Supongamos que, de un protón de un par, medimos la componente de su spin a lo largo del eje A, y cuyo valor sea A^+. No se realizan más mediciones de esta partícula; pero sí medimos, de su compañero del estado singlete, la componente de su spin a lo largo del eje B, cuyo valor resulte ser B^+. La última medición, que puede tomarse en un lugar distante después de que los protones se han ido alejando el uno del otro por cierto tiempo, nos da una información adicional acerca del estado del primer protón. Para ser explícitos, la existencia de una correlación negativa estricta implica que el primer protón, que ya sabemos a través de una medida directa que tiene la componente de spin A^+, debe tener también la componente B^-.

Esto significa que la observación de un par de protones, uno de los cuales tiene componente de spin A^+ y el otro componente de spin B^+ puede emplearse como una señal indicativa de la existencia de un único protón de componentes A^+B^-. Además, mediante un argumento estadístico puede probarse que $n[A^+B^+]$, el número de tales pares doblemente positivos, debe ser proporcional a $N(A^+B^-)$, número de protones individuales con las componentes de spin A^+B^-. Y asimismo $n[A^+C^+]$ debe resultar proporcional a $N(A^+C^-)$ y $n[B^+C^+]$ debe ser proporcional a $N(B^-C^+)$. La constante de proporcionalidad es, en los tres casos, la misma. Para protones individuales, sometido cada uno de ellos a una doble medición imaginaria, se ha demostrado una desigualdad, que afirma que $N(A^+B^-)$ no puede ser mayor que la suma de dos términos $N(A^+C^-)$ + $N(B^-C^+)$. Podemos sustituir cada una de estas cantidades no medibles por el correspondiente número de pares de protones doblemente positivos. La expresión resultante es

$$n[A^+B^+] \leq n[A^+C^+] + n[B^+C^+].$$

que constituye la desigualdad de Bell.

Obviamente, dicha desigualdad se prueba mediante ese razonamiento sólo si las tres premisas de las teorías realistas locales se consideran válidas. En efecto, aquí es donde las premisas tienen su aplicación más importante y, en último término, la más dudosa. De aceptarse las premisas, por la propia fuerza del razonamiento síguese que la desigualdad de Bell deberá satisfacerse. Más aún, nunca se especificó la orientación de los ejes A, B y C; y así, la desigualdad deberá ser

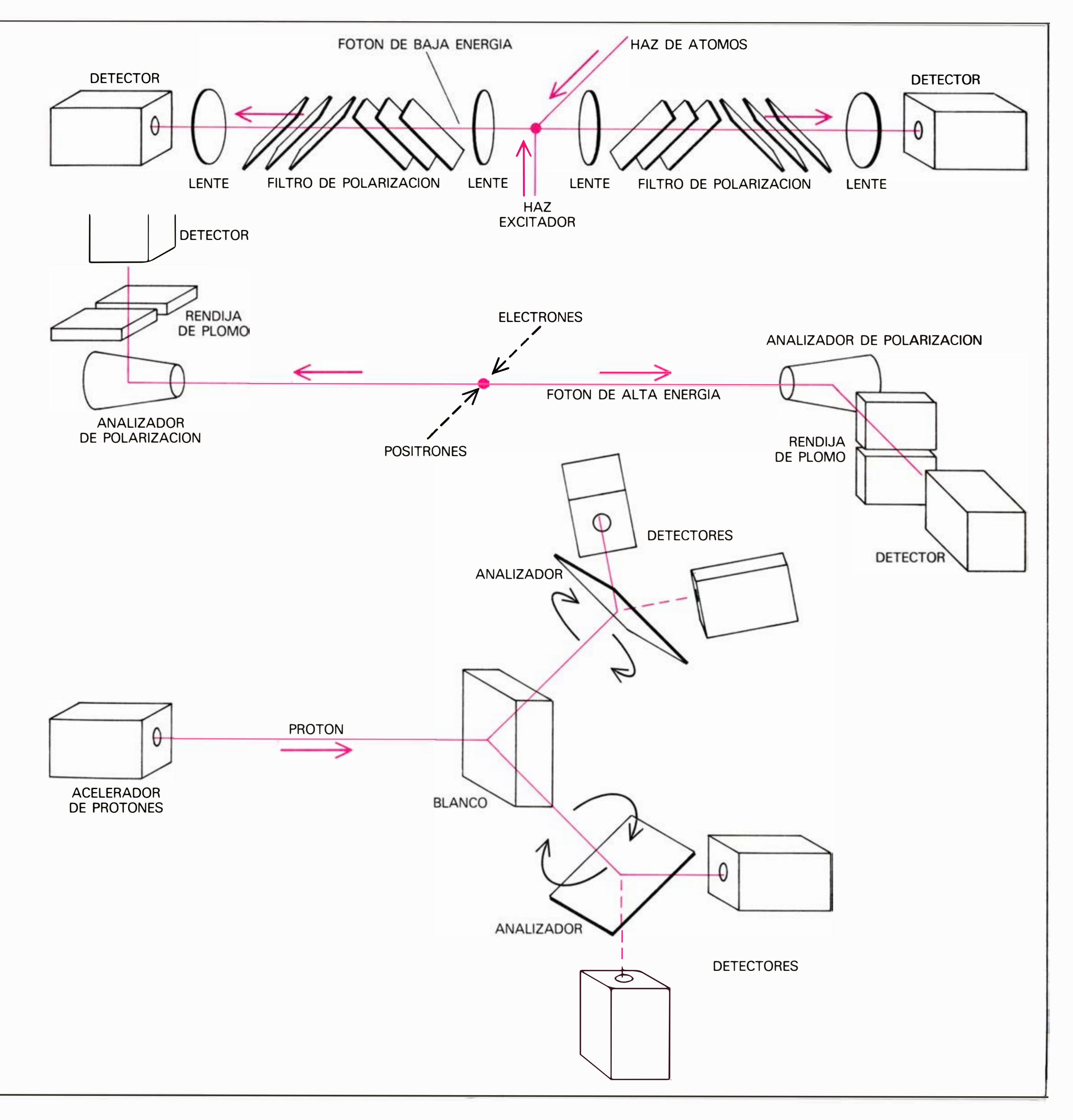
EXPERIMENTO	FECHA	PARTICULAS ESTUDIADAS	RESULTADOS
Stuard J. Freedman y John F. Clauser, Universidad de California en Berkeley	1972	Fotones de baja energía emitidos durante transiciones en átomos de calcio.	De acuerdo con la mecánica cuántica.
R. A. Holt y F. M. Pipkin, Universidad de Harvard	1973	Fotones de baja energía emitidos durante transiciones en átomos de mercurio 198.	De acuerdo con la desigualdad de Bell.
John F. Clauser, Universidad de California en Berkeley	1976	Fotones de baja energía emitidos durante transiciones en átomos de mercurio 202.	De acuerdo con la mecánica cuántica.
Edward S. Fry y Randall C. Thompson, Texas A & M University	1976	Fotones de baja energía emitidos durante transiciones en átomos de mercurio 200.	De acuerdo con la mecánica cuántica.
G. Faraci, S. Gutkowski, S. Notarrigo y A. R. Pennisi, Universidad de Catania	1974	Fotones de alta energía (rayos gamma) de la aniquilación de electrones y protones.	De acuerdo con la desigualdad de Bell.
L. Kasday, J. Ullman y C. S. Wu, Columbia University	1975	Fotones de alta energía (rayos gamma) de la aniquilación de electrones y positrones.	De acuerdo con la mecánica cuántica.
M. Lamehi-Rachti y W. Mittig, Centro de Investigaciones Nucleares de Saclay	1976	Pares de protones en el estado singlete.	De acuerdo con la mecánica cuántica.

PRUEBAS REALES de la desigualdad de Bell, realizadas por siete grupos de investigadores y la fecha respectiva de los mismos (*columna de la izquierda*). Sólo uno de los experimentos mide las componentes del spin de los protones; los otros estudian la polarización de los fotones, o cuantos de la radiación electromagnética. En cuatro de estos experimentos, se emplearon pares de fotones de baja energía con polarizaciones opuestas emitidos por átomos que se habían colocado en un estado excitado. En otros dos experimentos, se creaban pares de rayos gamma, o fotones de alta energía con polarizaciones opuestas al aniquilarse mutuamente electrones y sus antipartículas, los positrones. En el experimento restante, protones de un acelerador de partículas se hacían incidir sobre un blanco compuesto en parte por hidrógeno. Los protones acelerados y los núcleos de hidrógeno formaban pares en estado singlete. Cinco de los siete experimentos dieron resultados que violaban la desigualdad de Bell y que estaban de acuerdo con la mecánica cuántica. Los físicos admiten hoy que la desigualdad de Bell puede resultar violada. Seguimos desconociendo cuál es la causa de discrepancia de los dos experimentos restantes.

válida con independencia de los ejes elegidos. La única posible violación de la desigualdad podría ser resultado de una desviación estadística, donde muchas partículas con las componentes de spin A^+ y B^+ aparezcan con coincidencia fortuita. La probabilidad de esta coincidencia se aproxima a cero cuando el número de partículas medidas aumenta.

La desigualdad de Bell constituye una predicción explícita del resultado de un experimento. Las reglas de la mecánica cuántica pueden usarse para predecir los resultados del mismo experimento. No daré los detalles de cómo la predicción puede deducirse del formalismo matemático de la teoría cuántica; sin embargo, mencionaré que el procedimiento es totalmente explícito y objetivo, en el sentido de que cualquiera que aplique las reglas correctamente obtendrá el mismo resultado. Sorprendentemente, las predicciones de la mecánica cuántica difieren de las de las teorías realistas locales. En particular, la mecánica cuántica predice que, para algunas elecciones de los ejes A, B y C, se viola la desigualdad de Bell, de suerte que hay más pares de protones A^+ B^+ que pares combinados hay de A^+C^+ y B^+C^+. Por tanto, las teorías realistas locales y la mecánica cuántica son antagónicas.

El conflicto entre las mismas plantea dos cuestiones. En primer lugar, ¿cuáles son los hechos experimentales que dan



pie a esa situación? ¿Se satisface o se viola la desigualdad de Bell? Cualquiera que sea el resultado experimental, debe haber algún tipo de fallo en las reglas de la mecánica cuántica o en las teorías realistas locales. La segunda cuestión será, en consecuencia: ¿Qué premisa de la teoría refutada es falsa?

El experimento imaginario propuesto en 1935 por Einstein, Podolsky y Rosen suponía mediciones de la posición y momento de las partículas. El experimento sobre las componentes de spin del protón fue discutido por primera vez en 1952 por David Bohm, del Birkbeck College de Londres, aunque todavía en el contexto de un experimento imaginario. Hubo que esperar hasta 1969, después de que Bell hubiera introducido su desigualdad, para contemplar la posibilidad de experimentos reales que investigaran las cuestiones planteadas. La viabilidad de tales experimentos fue discutida por John F. Clauser, de la Universidad de California en Berkeley, R. A. Holt, de la Universidad de Western Ontario, y Michael A. Horne y Abner Shimony, de la Universidad de Boston. Hallaron que, para un experimento práctico, habría que generalizar de algún modo la desigualdad de Bell, y que todavía seguiría siendo posible una prueba significativa para confirmar las teorías alternativas.

No debemos pasar por alto las dificultades técnicas de los experimentos. En el experimento imaginario ambos protones de cada par alcanzan siempre los instrumentos y los instrumentos dan siempre una medición inequívoca de la componente del spin a lo largo del eje elegido. Los aparatos reales no pueden reproducir estos resultados. Los detectores no consiguen nunca un rendimiento perfecto: muchos protones pasan sin ser detectados. Debido a las imperfecciones de los instrumentos, el número de protones que se cuentan en cada categoría no puede interpretarse directamente. Hay que operar pues calculando con la ineficiencia de los contadores, imprecisión que se añade a la incertidumbre de los resultados.

De los siete experimentos realizados desde 1971, en seis no se medían las componentes del spin de los protones; en vez de ello, se medía la polarización de los fotones: los cuanta de la radiación electromagnética. La polarización es una propiedad del fotón que se corresponde con la del spin de una partícula material. En una serie de experimentos, se colocaban los átomos de un cierto elemento e isótopo en un estado excitado mediante la absorción de luz láser; y se les permitía luego volver al estado fundamental en dos pasos. En cada paso, se emitía un fotón de energía o longitud de onda característica. Los fotones se movían en direcciones opuestas y tenían polarizaciones también opuestas. En otras palabras, si se medía la polarización de ambos fotones a lo largo de una única dirección, se observaba una correlación negativa estricta.

En estos experimentos, la diferencia entre instrumentos ideales y reales es muy clara. No hay ningún aparato que pueda interceptar, por sí solo, un fotón y dar directamente su polarización. Necesitamos dos aparatos: un filtro y un detector. El filtro tiene por misión dejar pasar los fotones que detentan una polarización seleccionada y parar o desviar todos los demás. El detector cuenta el número de fotones que pasan a través del filtro. Ninguna de estas componentes es perfecta, de forma que un fracaso en el registro de un fotón no significa necesariamente que éste tuviera la polarización no deseada.

También se han realizado experimentos con rayos gamma, que son fotones de alta energía. Los rayos gamma se habían creado en el proceso de aniquilación mutua de electrones y de sus antipartículas, los positrones. Tal aniquilación da origen a dos rayos gamma, emitidos en direcciones opuestas y dotados de polarizaciones también opuestas. Los experimentos son, por tanto, formalmente equivalentes a los de física atómica, pero los aparatos que se precisan son muy distintos. En general, los detectores se muestran eficientes para fotones

RESULTADOS DE UNA PRUEBA EXPERIMENTAL donde se demuestra con toda claridad cómo se viola la desigualdad de Bell. Se trata del experimento que usaba pares de protones en el estado singlete, realizado por M. Lamehi-Rachti y W. Mittig, del Centro de Investigaciones Nucleares de Saclay, Francia. La correlación negativa entre los valores de distintas componentes del spin aparece representada en función del ángulo entre las direcciones de los dos analizadores. Una correlación de −1 indicaría que las componentes tienen invariablemente valores opuestos. La desigualdad de Bell establece que la correlación para cualquier ángulo debe hallarse sobre, o por encima de la línea coloreada. Las correlaciones observadas a 30, 45 y 60 grados están por debajo de esta línea. Los resultados violan la desigualdad de Bell y están en buen acuerdo con las predicciones de la mecánica cuántica, lo que les añade credibilidad. La violación de la desigualdad de Bell implica que al menos una de las tres premisas de las teorías realistas locales debe ser falsa. La separabilidad de Einstein (ninguna influencia puede propagarse más rápidamente que la luz) se considera como el candidato más plausible.

de alta energía, mientras que el mejor rendimiento de los filtros polarizadores lo obtenemos para los fotones de baja energía.

En un experimento se han medido las correlaciones de las componentes del spin de protones. Se parece mucho, pues, al experimento imaginario original. Los pares de protones se crean inyectando protones de energías relativamente bajas en un blanco constituido parcialmente por átomos de hidrógeno. El núcleo de un átomo de hidrógeno consta de un único protón. Cuando un protón incidente choca con un núcleo de hidrógeno, los dos protones interaccionan brevemente y forman un estado singlete. Ambos, a continuación, abandonan el blanco compartiendo el momento del protón incidente; si nada los perturbara, permanecerían en el estado singlete. Mediciones preliminares del mismo componente de spin de ambos protones dan resultados opuestos.

Los instrumentos para un experimento con pares de protones consisten, de nuevo, en filtros y detectores. En un experimento que se ha llevado a término, el filtro tenía una lámina de carbón, que dispersaba cada protón hacia uno de un par de detectores según fuera el valor de la componente medida.

Sin prestar atención a qué partículas se estudian, el experimento consta de tres series de dobles medidas. Se eligen tres ejes: *A*, *B*, y *C*. En general, los ángulos entre ellos se disponen de forma que correspondan a los valores donde se espera una mayor discrepancia entre la mecánica cuántica y las teorías realistas locales. Se coloca, entonces, un filtro que admita las partículas con la polarización o componente del spin A^+, y el otro se coloca de forma que deje pasar las partículas con componente B^+. Una vez registrada una muestra de partículas suficientemente grandes en esta configuración, se giran los filtros para medir las componentes a lo largo de los ejes *A* y *C*; se apuntan los nuevos datos. Por último se reorientan de nuevo los filtros para medir según los ejes *B* y *C*. A continuación, se cuentan las coincidencias registradas en cada configuración y se hacen las correcciones necesarias para cubrir la ineficiencia de los aparatos. Comparar los resultados con la desigualdad de Bell se reduce entonces a una simple suma.

De los siete experimentos terminados, cinco están de acuerdo con las predicciones de la mecánica cuántica. Es decir, señalan una violación de la desigualdad de Bell para algunas elecciones de los ejes *A*, *B* y *C*. Los otros dos dan correlaciones no mayores que las permitidas por la desigualdad de Bell y, por tanto, apoyan las teorías realistas locales. El tanteo es de cinco a dos en favor de la mecánica cuántica. En realidad, el apoyo en favor de la mecánica cuántica es mucho más fuerte de lo que dicha relación parece implicar. Así, una razón para atribuir mayor credibilidad a los cinco experimentos que violan la desigualdad de Bell es que éstos representan una muestra de datos mayor y, por tanto, son estadísticamente más significativos. Algunos de estos experimentos se acometieron después de que se hicieran públicos los dos resultados anómalos; e incorporaron refinamientos en la instrumentación diseñados explícitamente para evitar toda desviación que pudiera ser el origen de los dos resultados discrepantes. Clauser y Shimony han hecho notar que también hay una justificación epistemológica para no tener en cuenta los dos experimentos que están en desacuerdo con la mayoría. La mecánica cuántica predice una mayor correlación entre los sucesos y, una menor correlación entre ellos, las teorías realistas locales.

Gran variedad de fallos sistemáticos en un experimento podrían destruir la evidencia de una correlación real, obteniendo así resultados dentro de los límites impuestos por la desigualdad de Bell. Por otra parte, es difícil imaginar un error experimental que pudiera crear una correlación falsa en cinco experimentos independientes. Más aún, los resultados de estos experimentos, además de violar la desigualdad de Bell, lo hacen precisamente de la forma predicha por la mecánica cuántica. Para que los resultados de los cinco experimentos se debieran a coincidencias fortuitas se exigiría una desviación estadística extraordinaria, increíble habida cuenta del número de partículas detectadas ahora.

Se están pensando ya en nuevas pruebas de la desigualdad de Bell. Y hay en preparación otro nuevo experimento, por lo menos. La mayoría de los físicos ocupados en estos problemas tienen la seguridad radical, fundada en los cinco resultados coherentes, de que el problema ha sido resuelto. Para algunas elecciones de los ejes *A*, *B* y *C*, la desigualdad de Bell se viola en la naturaleza; por consiguiente, las teorías realistas locales son falsas.

Si podemos establecer que se ha demostrado la falsedad de las teorías realistas locales, ¿qué premisa fundamentante de las mismas es la falsa? A la hora de contestar la pregunta, como primer paso, convendría estar seguros de que no se han hecho hipótesis adicionales al formular la prueba experimental.

Pero sucede que se requirió una hipótesis subsidiaria, al menos. Debido a las limitaciones de los instrumentos prácticos, hubo que generalizar ligeramente la desigualdad de Bell, generalización que se tuvo que aceptar por válida. No puede probarse. Parece muy improbable, sin embargo, que esta circunstancia llegara a alterar los fenómenos de suerte que los resultados de los experimentos no sólo violaran la desigualdad de Bell sino que, además, se mostraran concordes con las predicciones de la mecánica cuántica. En todo caso, cabe esperar que experimentos más refinados prueben la desigualdad de Bell sin la generalización. Como la hipótesis subsidiaria es susceptible de una comprobación experimental, parece menos fundamental que las otras tres, y, por tanto, no la consideraremos en lo que sigue.

Otro campo que pudiera analizarse para hipótesis no reconocidas es la prueba de la desigualdad de Bell. En efecto, todo indica que la prueba depende de la supuesta validez de la lógica ordinaria bivalente, en la que toda proposición debe ser verdadera o falsa y una componente del spin debe ser o más o menos. Algunas interpretaciones de la mecánica cuántica han introducido la idea de una lógica plurivalente, pero esas consideraciones no tienen nada que ver con los razonamientos aplicados en esta prueba. En efecto, en el contexto de la prueba es difícil hasta imaginar una alternativa a la lógica bivalente. Mientras no se formule un tal sistema, lo mejor será olvidarse de este problema.

El bloque entero de experimentos fundados en las ideas de Einstein, Podolsky y Rosen se mira a veces como una mera prueba de las teorías de variables ocultas. Los experimentos comprueban realmente esas teorías, pero debe hacerse hincapié en que la existencia de variables ocultas no es ninguna nueva premisa de las teorías locales. Por el contrario, la existencia de parámetros que especifiquen las propiedades deterministas de una partícula se dedujo a partir de las tres hipótesis originales. Recuérdese que el psicólogo no supuso que la prueba que había inventado midiera ningún atributo real de los individuos sometidos al test; antes bien, dedujo la existencia de tal atributo al observar una correlación estricta. De igual modo, se postuló la existencia de variables ocultas a raíz de la correlación negativa que se detectó al medir una sola componente

del espín en pares de protones en estado singlete.

Quizá no pueda probarse con rigor que, en la argumentación en pro de las teorías realistas locales, no intervenga ninguna otra hipótesis suplementaria. De todos modos, la cadena del razonamiento es lo suficientemente sencilla como para suponer que si se hallaran implícitas algunas otras hipótesis, serían fácilmente reconocibles. No se ha encontrado todavía ninguna. Centraremos, pues, nuestra atención en las tres premisas: realismo, libre uso de la inducción y separabilidad de Einstein.

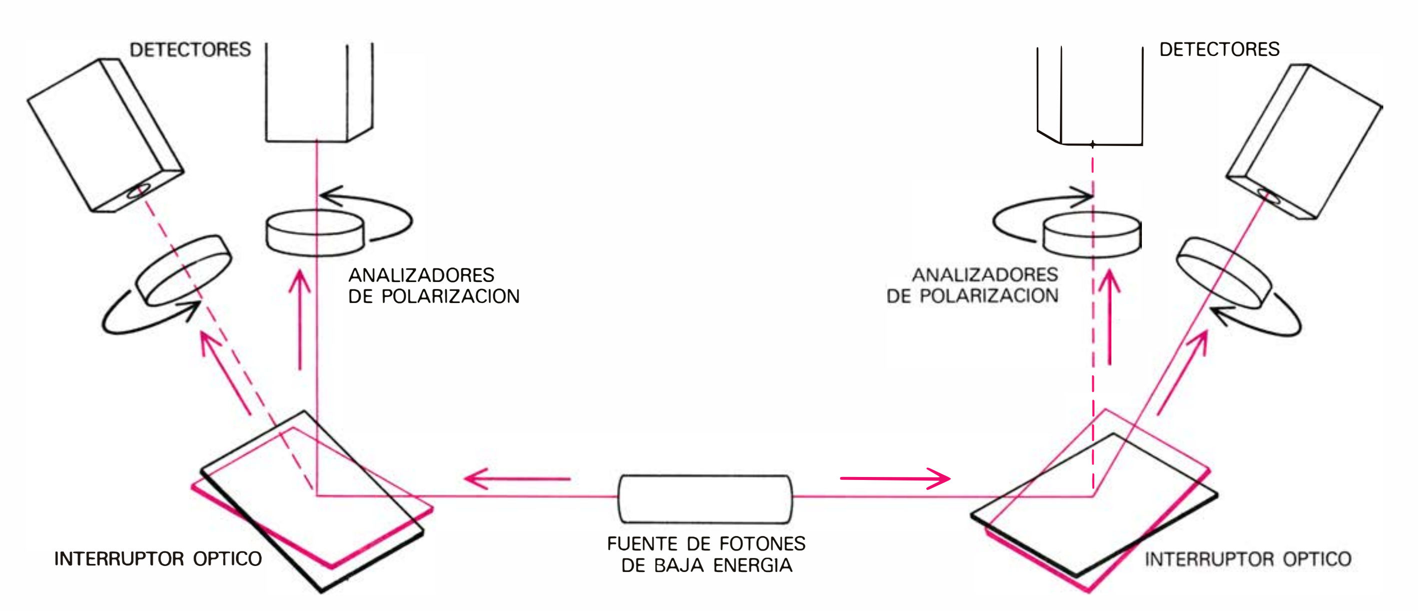
De las tres, el realismo constituye la premisa fundamental. Puede enunciarse formalmente así: debemos exigir a una teoría algo más que una mera descripción de los datos. Ni tan siquiera basta una regla empírica para predecir los resultados de futuros experimentos. La mente pide algo más: no necesariamente determinismo –no hay nada intrínsecamente irracional en el carácter probabilístico–, sino, por lo menos, una explicación objetiva de las regularidades observadas, o en otras palabras causas. Bajo esta exigencia subyace la noción intuitiva de que el mundo exterior a nosotros es real y que tiene al menos algunas propiedades que existen independientemente de la conciencia humana.

Cierto número de filósofos, que podemos englobarlos bajo el calificativo de positivistas, han rechazado el punto de vista realista. Los positivistas no niegan la existencia del mundo exterior a la mente; llanamente, consideran sin sentido cualquier afirmación acerca de la realidad externa que no se refiera de un modo directo a las impresiones sensoriales. En el siglo XX algunos positivistas radicales han tenido una influencia apreciable, bien que indirecta, sobre la manera de pensar de la física teórica.

El sentido de paradoja inducido por el hallazgo de la violación de la desigualdad de Bell puede mitigarse adoptando una actitud positivista. Camino que fue propuesto hace ya tiempo. Cuando se piensa en todas las consecuencias que comportaría el abandono del realismo se descubre, empero, que sería una renuncia demasiado grande para que resulte atrayente. En el contexto de esta experiencia, el positivismo afirma que no tendría sentido atribuir nada que se pareciera a una componente de spin bien definida de una partícula antes que medir dicha componente; que la única magnitud con una realidad verificable es la misma observación, la impresión sensorial; y que debe rechazarse, en última instancia, la exigencia del psicólogo de una explicación objetiva de la notable correlación que halla. Si se aplicara de un modo coherente esa negativa a buscar causas subyacentes de las regularidades observadas, la ciencia quedaría enteramente reducida a una trivialidad. La ciencia quedaría reducida a un recetario para predecir las observaciones futuras a partir de las ya realizadas. Cualquier noción de ciencia como "el estudio de la naturaleza" sería imposible. La naturaleza pasaría a ser pura ilusión. Podemos imaginar una física basada en principios positivistas, capaz de predecir todas las correlaciones de sucesos y dejar todavía el mundo totalmente incomprensible. Dadas las consecuencias extremas a que nos llevaría la abolición del realismo, uno se inclina por aferrarse a la primera premisa.

En la argumentación en que se basan las teorías realistas locales, el realismo aparece unido a otro punto: se trata de la justificación de postular el uso libre de la inducción. La inducción ha permitido al físico extrapolar, a partir de una serie de correlaciones negativas observadas, la conclusión de que cualquier par de protones en el estado singlete tiene valores opuestos de una cualquiera de las componentes del spin, aun cuando no se mida ninguna de las componentes. La extrapolación constituyó un paso esencial en la prueba de la desigualdad de Bell; pero es indefendible si el concepto de propiedades no medidas carece de sentido.

Este uso de la inducción pudiera parecerles a algunos como el eslabón débil en la cadena de la argumentación. Poco después de que apareciera el trabajo de Einstein, Podolsky y Rosen, Niels Bohr publicó una réplica en la que defendía la completitud de la descripción mecánico-cuántica de la naturaleza. Fundaba su



SEPARABILIDAD DE EINSTEIN, cuya comprobación rigurosa se espera ahora del experimento que está preparando Alain Aspect, del Instituto de Optica de París. Experimentos anteriores sólo han corroborado el principio de separabilidad, que es menos restrictivo: las posiciones de los analizadores se determinaban mucho antes de las mediciones, de forma que cierta influencia de una medición podría comunicarse (a través de un mecanismo desconocido) a la otra medición a una velocidad inferior a la de la luz. Esta posible explicación de la correlación observada resulta improbable en grado sumo, pero quedaría totalmente excluida si la posición de los analizadores cambiara tan rápidamente que ninguna señal que no se moviera más rápidamente que la luz pasara de un detector a otro con tiempo para condicionar el resultado de la segunda medida. En el experimento de Aspect, que medirá polarizaciones de fotones de baja energía, esta condición se cumplirá. Por cada fotón habrá dos conjuntos de analizadores y detectores, y los analizadores medirán componentes distintas. Un interruptor óptico rápido determinará en qué analizador entra el fotón sólo cuando ya sea tarde para que esta decisión pueda influir en la otra medida. (Suponiendo que la influencia se propaga a velocidad menor que la luz.) El interruptor se representa como un espejo móvil. El efecto deseado se logrará mediante ondas de ultrasonidos que incidirán en la superficie de un cristal.

crítica en que el uso de la inducción que hacía Einstein no estaba justificado. La contestación de Bohr constituye un documento central de lo que posteriormente se dio en llamar interpretación de Copenhague de la mecánica cuántica. Su razonamiento pudiera sintetizarse así: una partícula y un instrumento preparado para tomar una medición específica de la misma forman de alguna manera un solo sistema, que quedaría alterado de un modo esencial si se cambiara el dispositivo instrumental. Por esta causa no está permitido hacer inferencias sobre el estado de la partícula sin especificar al mismo tiempo las posiciones de los instrumentos que interaccionarán con la partícula.

Los puntos de vista de Bohr han ejercido una amplia influencia en muchos físicos. En un cierto sentido, esto es bueno. Después de todo, los trabajos recientes que estamos discutiendo han demostrado que en estas materias él estaba más cerca de la verdad que Einstein. Sin embargo, cuando examinamos las ideas de Bohr en su estricto contenido, surgen objeciones muy parecidas a las que se levantaron contra la postura positivista. Porque el realismo nos da la última causa racional para el uso libre de la inducción, podemos argüir que Bohr no era realista, o no lo era al menos de una manera coherente. Cualquier explicación de los experimentos de correlación a distancia que se base en la réplica de Bohr a Einstein, Podolsky y Rosen puede resultar inconsistente incluso con una versión moderada del realismo.

Si hemos de mantener el realismo y el uso libre de la inducción, la violación de la desigualdad de Bell podrá sólo explicarse en la suposición de que no sea válida la hipótesis de separabilidad de Einstein. En el experimento del psicólogo, se entendía la separabilidad en el sentido de que los maridos y las mujeres, una vez divididos, no podían comunicarse entre sí. En el experimento físico, la hipótesis de separabilidad expresaba la idea, intuitivamente razonable, de que las componentes de spin de un protón no influían en las del otro protón, si las dos partículas se hallaban suficientemente alejadas. La hipótesis más restrictiva de la separabilidad de Einstein prohíbe tal influencia sólo si se propagara con una velocidad mayor que la de la luz. Como he probado, esta hipótesis debe tomarse ahora como altamente cuestionable.

Antes de afrontar el estudio de las consecuencias de esta conclusión debe hacerse notar que ninguno de los experimentos hasta ahora realizados ha

corroborado, en rigor, la hipótesis de separabilidad de Einstein. En otros experimentos, las posiciones de los instrumentos quedaban determinados mucho antes (en la escala de tiempos de la física de partículas). Por tanto, la disposición de un instrumento podría, razonablemente, afectar sucesos observados en el otro instrumento, o podría modificar variables ocultas en la fuente de pares de protones; en ningún caso de ambos se precisaría que la influencia viajara más de prisa que la luz. Un experimento con instrumentos cuyas posiciones cambiaran rápidamente eliminaría esta posibilidad. La decisión de medir cierta componente del spin con un detector no se haría hasta que fuera demasiado tarde para que cualquier influencia de esta decisión pudiera alcanzar el otro instrumento o la fuente, incluso a la velocidad de la luz, con tiempo para alterar el resultado de la segunda medición. Tal experimento se está realizando por Alain Aspect, del Instituto de Optica de la Universidad de París.

Independientemente del problema sobre la rapidez con que pueda viajar una influencia hipotética de un instrumento al otro, la existencia misma de esta influencia parece muy poco probable. Habría de alterar las observaciones distantes y, ello, de la manera necesaria para producir la violación observada en la desigualdad de Bell. Parece pues más indicado buscar otra explicación y suponer, mientras esperamos los resultados del experimento de Aspect, que si se viola la separabilidad ordinaria, lo mismo le ocurrirá a la separabilidad de Einstein.

A lo largo del artículo he venido considerando el par de protones como si se tratara de entidades independientes que se reunieran en el blanco y luego volvieran a disgregarse. Pero también pueden entenderse como elementos de un sistema físico único que se crea durante la primera interacción y progresivamente se va extendiendo más y más en el espacio hasta que la primera medición lo destruye. Por lo que respecta a la separabilidad, ambas explicaciones son equivalentes. En cada caso, una violación de la separabilidad de Einstein requiere acción a distancia instantánea, ya sea entre sistemas independientes, ya sea dentro de un único sistema desplegado.

¿Debe, pues abandonarse el principio de propagación con velocidad finita de las señales? No hay que responder la cuestión de un modo apresurado. El principio se introdujo como una premisa de la teoría de la relatividad, y sin él ésta pierde su coherencia intrínseca. Más aún, señales que viajen más rápidamente que la luz originarán paradojas extrañas de causalidad: los observadores de algunos sistemas de referencia hallarán que un suceso está "causado" por otro que aún no ha tenido lugar. Sin embargo, las influencias instantáneas que parece han de operar en los experimentos de correlación a distancia no exigen una revisión tan drástica de las ideas en boga. Resulta bastante verosímil que tales influencias no podrían emplearse para transmitir ninguna información "útil", órdenes o instrucciones por ejemplo. Ningún suceso que ocasiona otro suceso puede ligarse al segundo mediante este mecanismo; las influencias instantáneas pueden transmitirse sólo entre sucesos que están relacionados por una causa común. Por tanto, el concepto de señal tendría que volver a definirse en el sentido de que sólo aquellos medios de comunicación que transmitan información útl deberían denominarse señales. Y el principio de velocidad finita para las señales quedaría a salvo.

No obstante, incluso esta solución llega a poner en cierto peligro el realismo científico. La ley fundamental de que las señales no pueden viajar a velocidades superiores a la de la luz ve menguada su importancia; de constituir una propiedad de la realidad externa pasa a ser una mera característica de la experiencia humana comunicable. Aunque ésta representa dar un paso hacia el positivismo filosófico, el concepto de una realidad independiente o externa puede seguir defendiéndose como una explicación posible de las regularidades observadas en los experimentos. Sin embargo, es necesario que la violación de la separabilidad de Einstein quede incluida como una propiedad, aunque una propiedad bien escondida y contra la intuición, de esta realidad independiente. Debemos mencionar la pasada que la refutación de Bohr del argumento de Einstein, a propósito de la existencia de variables ocultas, introduce una violación implícita de la separabilidad. Se funda en una extraña indivisibilidad del sistema de partículas y de los instrumentos de observación.

La argumentación que procede desde una correlación observada entre la desigualdad de Bell y la violación de la separabilidad de Einstein no es particularmente complicada, pero sí indirecta. ¿Podría haberse obtenido el mismo resultado de una forma más directa? Sucede que no podría haberse demostrado sin la desigualdad de Bell, aunque si pudo sospecharse, cosa que ocurrió. La sospecha se funda en que la función de ondas para un sistema de dos o más partículas suele ser una entidad no local, de la que se supone que se colapsa repentinamente e incluso instantáneamente al verificar una medición. Si nos imaginamos la función de ondas como una gelatina real de un tipo especial, el colapso instantáneo viola claramente la separabilidad de Einstein. Pero ese cándido supuesto no se tomó nunca en serio, porque la interpretación convencional de la mecánica cuántica no identifica la función de ondas de un sistema con nada que pueda entenderse como la realidad del sistema. Bohr, por ejemplo, consideraba la función de ondas como mera herramienta para calcular. Además, describe la función de ondas para un sistema de varias partículas sólo en un enfoque que ignora la teoría de la relatividad; por tanto, su estructura difícilmente puede considerarse un argumento convincente contra la separabilidad de Einstein. Debido a estas razones, se podía creer, hasta hace unos pocos años, en una realidad externa, independiente y, al propio tiempo, considerar la separabilidad de Einstein como una ley completamente general aplicable a dicha realidad.

Una respuesta razonable a los experimentos de correlación a distancia es que su resultado no tiene consecuencias. Los mismos experimentos podrían representar una rara, y por tanto interesante, prueba de los fenómenos mecánico-cuánticos observados a gran distancia; pero los resultados no dan más de lo que se esperaba. Demuestran que la teoría está de acuerdo con la experimentación y, por tanto, no suministran información nueva. Esta reacción sería muy superficial. Verdad es que los experimentos, ahora que ya se han llevado a cabo, han resultado tener poco que ver con la mecánica cuántica. Pero eso no los trivializa, sino que indica que su importancia real está en otro lugar. Un descubrimiento que desacredita una hipótesis básica acerca de la estructura del mundo, una hipótesis mantenida desde hace tiempo y raramente puesta en duda, no es, evidentemente, trivial. Se trata de una iluminación que merece el reconocimiento.

Muchas partículas o agregados de partículas, que se les suele considerar objetos separados, han interaccionado con otros objetos, en algún momento del pasado. La violación de la separabilidad implica que, en algún sentido, todos esos objetos constituyen un todo indivisible. Quizás en un mundo así la idea de una realidad con existencia independiente pueda mantener parte de su significado, pero será un significado alterado y alejado de la experiencia ordinaria.



Una estación neolítica y de la Edad de Hierro en una colina inglesa

Los celtas con que tropezaron los romanos cuando éstos invadieron Britannia habían construido fortificaciones en las colinas. En Crickley, una de ellas se superpone a enigmáticas obras que precedieron a los celtas en 2000 años

P. W. Dixon

Cuando los romanos invadieron Britannia en el 43 d. C. tropezaron con la floreciente sociedad de los celtas. Los monumentos más característicos e impresionantes que los celtas han dejado son sus imponentes fortificaciones en lo alto de las colinas, de una extensión de varios centenares de hectáreas. La mayoría de estas fortificaciones están señaladas hoy día por los arruinados restos, cubiertos de hierba, de terraplenes y fosos defensivos. Las más primitivas de ellas se atribuyen por lo general a finales del segundo milenio a. C., mientras que las más recientes están consideradas como de fines del primer milenio a. C. Cualquiera que fuese la función de estas colinas fortificadas –"castros" en la terminología más usual en España– entre las comunidades de las que nacieron, su construcción supone una inversión tan inmensa de destreza, trabajo y materias primas, que un análisis de los castros es esencial para la comprensión de la estructura y del desarrollo cultural y económico de la sociedad en las postrimerías de la Europa prehistórica. Los castros indican claramente que además de las ocupaciones pacíficas de la época –agricultura, prácticas religiosas, artesanía y comercio– había un componente substancial de actividad emigratoria y militar.

En el sur de Inglaterra se conocen los restos de más de mil castros. Los trabajos de campo no destructivos –prospección de alrededores, recogida de hallazgos casuales levantados por la agricultura o la erosión, medición de anomalías del terreno con instrumentos geofísicos– han proporcionado información sobre la posición, tamaño y forma de muchas fortificaciones, pero los datos de superficie de montículos cubiertos de hierba no bastan para reflejar en el plano el desarrollo de las estaciones más complejas. Por ello, los arquéologos han invertido mucho trabajo en la excavación de cientos de castros, con distintos grados de acierto.

El estudio de los castros se ha resentido de la comprensible tendencia de los excavadores a concentrarse en los baluartes erguidos, especialmente en las entradas visibles, donde una modesta labor puede proporcionar una substanciosa cantidad de información acerca de cómo el edificio evolucionó. Una tendencia pareja ha sido la de prestar atención a las estaciones más grandes y más fuertes, las que cuentan con probabilidades de haber sido las de mayor significación, cuando menos en la política de la época. Sin embargo, aun la mayor de las excavaciones no puede abarcar de modo adecuado una superficie de quince a veinte hectáreas, a lo que se añade que el terreno cubierto de hierba o de maleza silvestre del interior de los castros del sur de Inglaterra, aparece a los ojos del arqueólogo que realiza la prospección como un lugar ingrato para excavar. Como consecuencia, es mucho lo que se sabe acerca de las defensas de los castros y poco acerca de los poblados que éstos encierran. Lo que se necesitaba era un castro de reducido tamaño, cuya excavación total y a mano pudiera ser llevada a cabo en el término de la vida de un sólo director, y cuyos restos no estuviesen afectados por cultivos o por anteriores excavaciones. En 1968 encontré una estación de este tipo en los Cotswolds de Gloucestershire, dominando el valle del Severn, a siete kilómetros al sur de Cheltenham y a nueve al este de Gloucester. Se la llama Crickley Hill.

La abrupta ladera occidental del banco de caliza de Cotswold está dividida por pequeños valles en colinas de cima llana y de forma de espolón. Muchas de estas cimas se hallaban fortificadas mediante terraplenes y fosos. El más exterior de los bastiones de Crickley Hill, de 300 metros de largo, delimita un área triangular de unas tres hectáreas. Durante todos los veranos desde 1969 excavadores y supervisores voluntarios han venido excavando, bajo mi dirección, casi la mitad de la cima de la colina. Nuestros trabajos han puesto al descubierto signos de actividad prehistórica de inesperada complejidad y duración. Hemos descubierto que Crickley Hill fue la estación, no sólo de dos castros celtas sucesivos, de la Edad de Hierro, sino también de dos recintos neolíticos pre-celtas.

La cima de Crickley Hill no es llana del todo; en el centro del castro se alza una loma pequeña que según nos indicó la fotografía aérea, estaba rodeada de un terraplén y un foso borrados por la erosión. En esta zona encontramos cinco fases sucesivas de ocupación neolítica. Los restos del edificio más antiguo y más pequeño se hallaban en lo más alto de la loma. Aquí, una serie de fosos pequeños formaba un cerco aproximadamente ovalado, de unos ocho metros de largo. El relleno de los fosos había sido apilado formando un túmulo en el centro: los intervalos que separaban los fosos habían sido desnudados de hierba y de tierra vegetal, dejando al descubierto el lecho de roca. Ignoramos por qué se hizo esto, pero la obra acabada debió parecerse a un diminuto túmulo

VISTA AEREA de la zona central del poblado de la Edad de Hierro de Crickley Hill, mostrando las señales de dos distintas fases de ocupación. El anillo de hoyos de postes excavado cerca del centro sostenía los soportes de la casa redonda mayor del segundo castro. Son visibles también los hoyos de los postes de las casas del castro más antiguo. Los amontonamientos paralelos de turba en lo alto del área excavada señalan la entrada de los dos castros. Abajo, a la derecha, descansan algunos voluntarios.

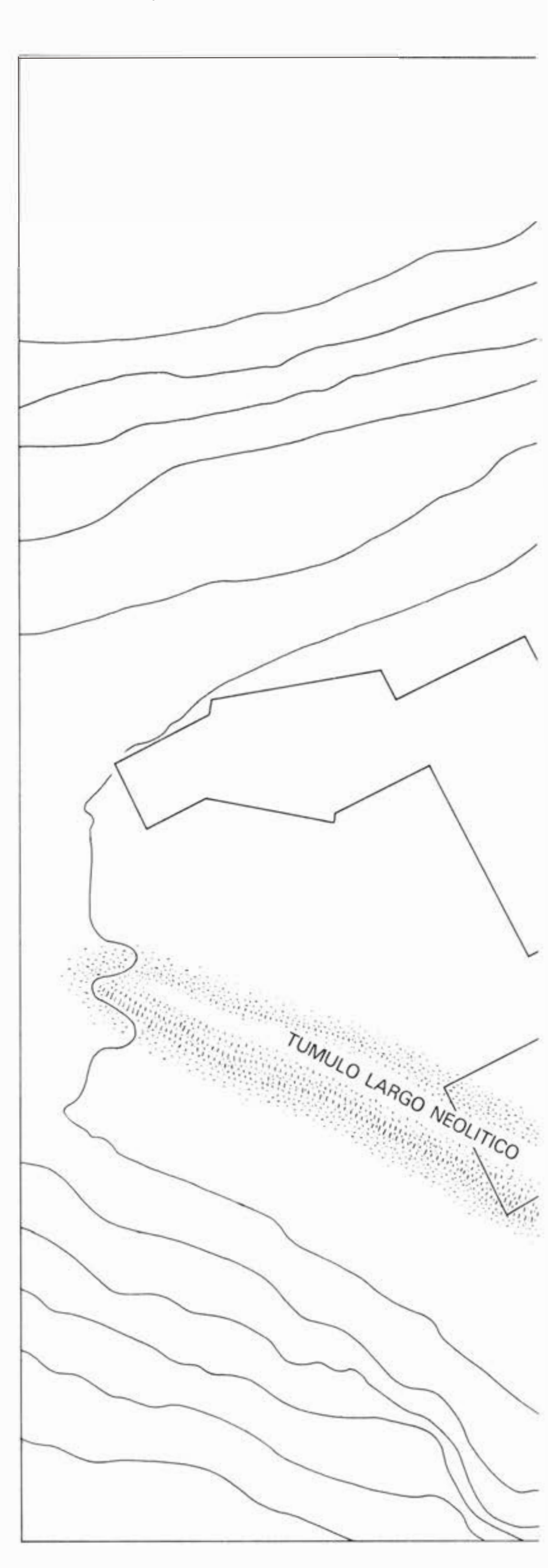
funerario. Posteriormente, la mayor parte del túmulo fue retirada y los fosos rellenados con material limpio; nuestra propia excavación extrajo de ellos casi diez toneladas de piedras y tierra (pero sólo exhumó un trozo pequeño de hueso). Esta construcción fue luego enterrada bajo el terraplén del primer recinto neolítico; una doble línea de fosos cortó el espolón triangular de la colina, de cima plana.

En el sur de Inglaterra se conocen por excavaciones, o se barrunta su existencia a través de fotografías aéreas, unos cincuenta recintos neolíticos. Se les ha dado en llamar "campamentos de calzadas" (*causewayed camps*) porque los fosos que los rodean están cruzados por muchas calzadas, y en general constan de dos o más anillos concéntricos. Cuando se descubrieron por primera vez a comienzos del presente siglo, se creyó que eran defensas de poblados; también se les consideró como ciudades neolíticas. Las excavaciones de la década de 1930 en Windmill Hill (Wiltshire) hicieron surgir las dudas, al no encontrarse restos de edificios dentro del recinto. Aunque la erosión del yeso del subsuelo pudiera haber borrado todas las huellas de las construcciones, la mayoría de los arqueólogos llegó a aceptar la opinión de que los recintos no eran defensivos, por aquella razón y por otras dos más: (1), porque los recintos parecían tener numerosas entradas, y (2), porque los constructores neolíticos desdeñaban con frecuencia el lugar de mejor defensa en beneficio de otro que no era tan fuerte.

A consecuencia de ello, se generalizó la opinión de que los campamentos eran probablemente corrales para el ganado, utilizados quizá para los rodeos del final del verano. El estudio de los huesos de animales que se han exhumado en algunas de las estaciones acusa el predominio de los huesos de ganado joven, pero no señala los de las terneras que uno esperaría encontrar como resultado de una campaña de matanza de invierno. Algún tipo de muerte ceremonial es lo que ahora se considera probable. Los enterramientos de cráneos y de patas de las bestias en posiciones claras en los fosos indican que los recintos eran los centros espirituales de las comunidades que los construyeron. Recientemente, en Hambledon Hill (Dorsetshire), las excavaciones descubrieron cráneos humanos colocados en posiciones idénticas, a intervalos regulares, en el fondo del foso de circunvalación, lo que el excavador interpreta como señal de que el recinto era una gran necrópolis de los poblados vecinos. Cuanto más trabajo se realiza, más se complica el problema, y la opinión actual se inclina a considerar que los campamentos de calzadas servían para una multitud de funciones prácticas y rituales, algunas de ellas de múltiples destinos, y otras especializadas. A decir verdad, los diversos campamentos de calzadas han podido tener tan sólo en común la interrupción de sus fosos por calzadas.

La importancia de las calzadas quizá haya sido exagerada en el pasado, porque los fosos han proporcionado el bloque de datos principal, mientras que los daños hechos por la erosión y por el arado han hecho desaparecer los terraplenes que los acompañaban y los niveles formados por la ocupación humana. En Crickley Hill los terraplenes están relativamente bien conservados y proporcionan un cuadro detallado de cómo estaba construido el recinto. Las dos líneas de fosos están cruzadas por 18 calzadas, pero 14 de éstas carecían de los huecos correspondientes en el terraplén del fondo y nunca dieron acceso al recinto. Aun así, el número de entradas sigue siendo grande para una fortificación defensiva: cuatro en 140 metros, con un número original que es probable haya sido de seis para 260 metros. Los agujeros para postes en los pasos de entrada indican, pese a todo, que los accesos se cerraban mediante puertas bien seguras. La puerta central del anillo interior da pruebas de una cuidadosa planificación: después de excavar el foso hasta una profundidad de unos 30 centímetros, la obra se detuvo en ella, y un tramo de la misma se rellenó para hacer la calzada de acceso rematada por una pavimentación de carretera. Basuras y restos de ocupaciones posteriores cubrieron la carretera y se acumularon en el fondo del foso. El hecho de que la abertura de la puerta central se corresponda con otra del foso exterior indica que ambos fosos, el de fuera y el de dentro, eran parte del mismo plan y que ese plan fue modificado, a propósito o por distracción, durante la construcción del recinto.

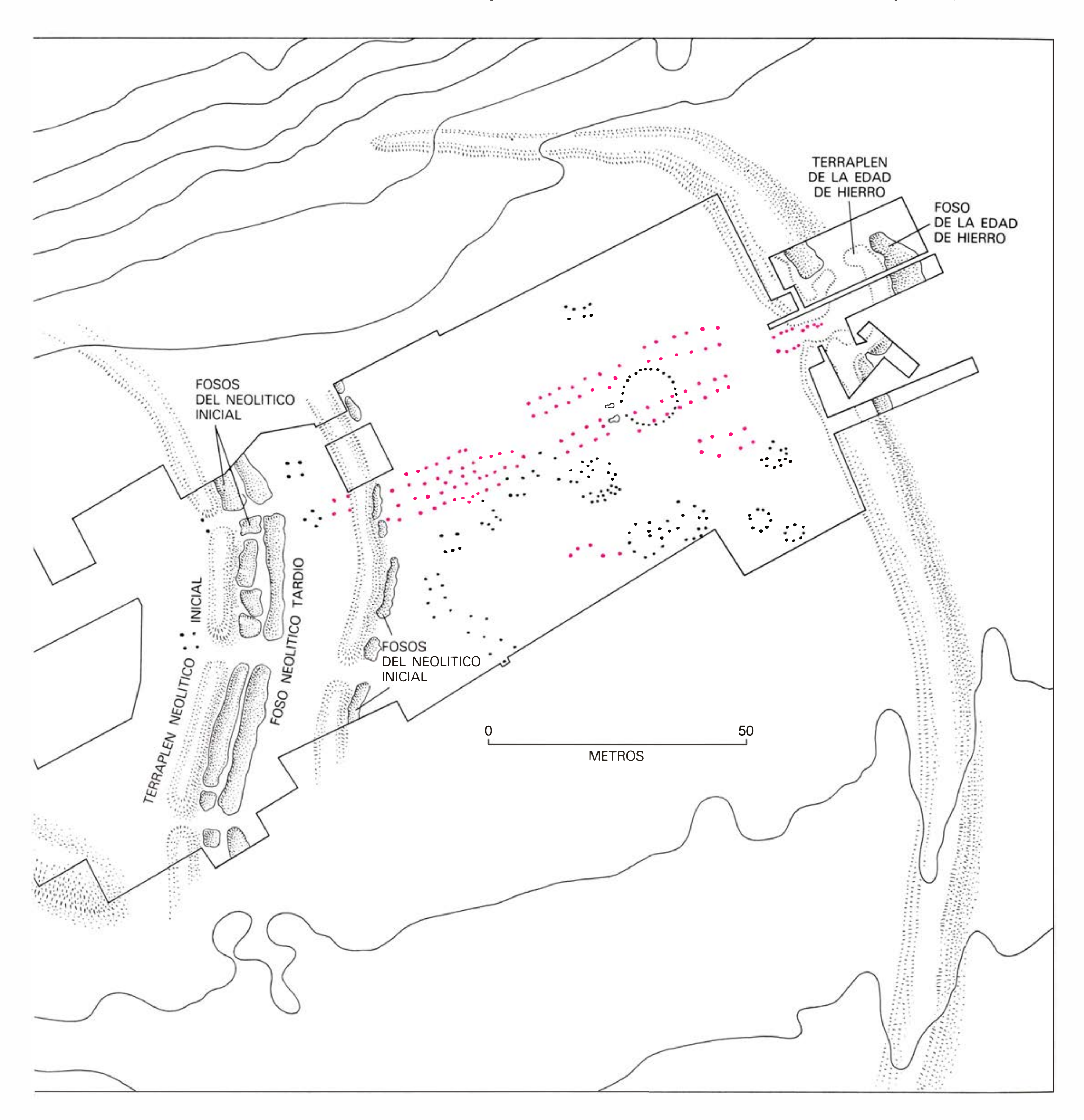
Las excavaciones en lo alto de la loma del centro de Crickley Hill han descubierto cientos de agujeros prehistóricos para postes, que se rellenaron, y lo mismo atarjeas y fosas pequeñas. A grosso modo pueden establecerse diferencias entre los rasgos de la Edad de Hierro y los del Neolítico, pero es mucho más difícil distinguir las varias fases de la ocupación neolítica. Según los estudios de la cerámica asociada con los niveles neolíticos, todas las fases del Neolítico tienen en común una cultura tan similar, que la separación de sus elementos culturales en períodos carecería de garantía. Aún no se dispone de fechas de carbono-14 para estas fases primitivas, pero la cerámica sugiere una duración para los recintos neolíticos de entre 3500 y 2500 a. C. Hasta ahora



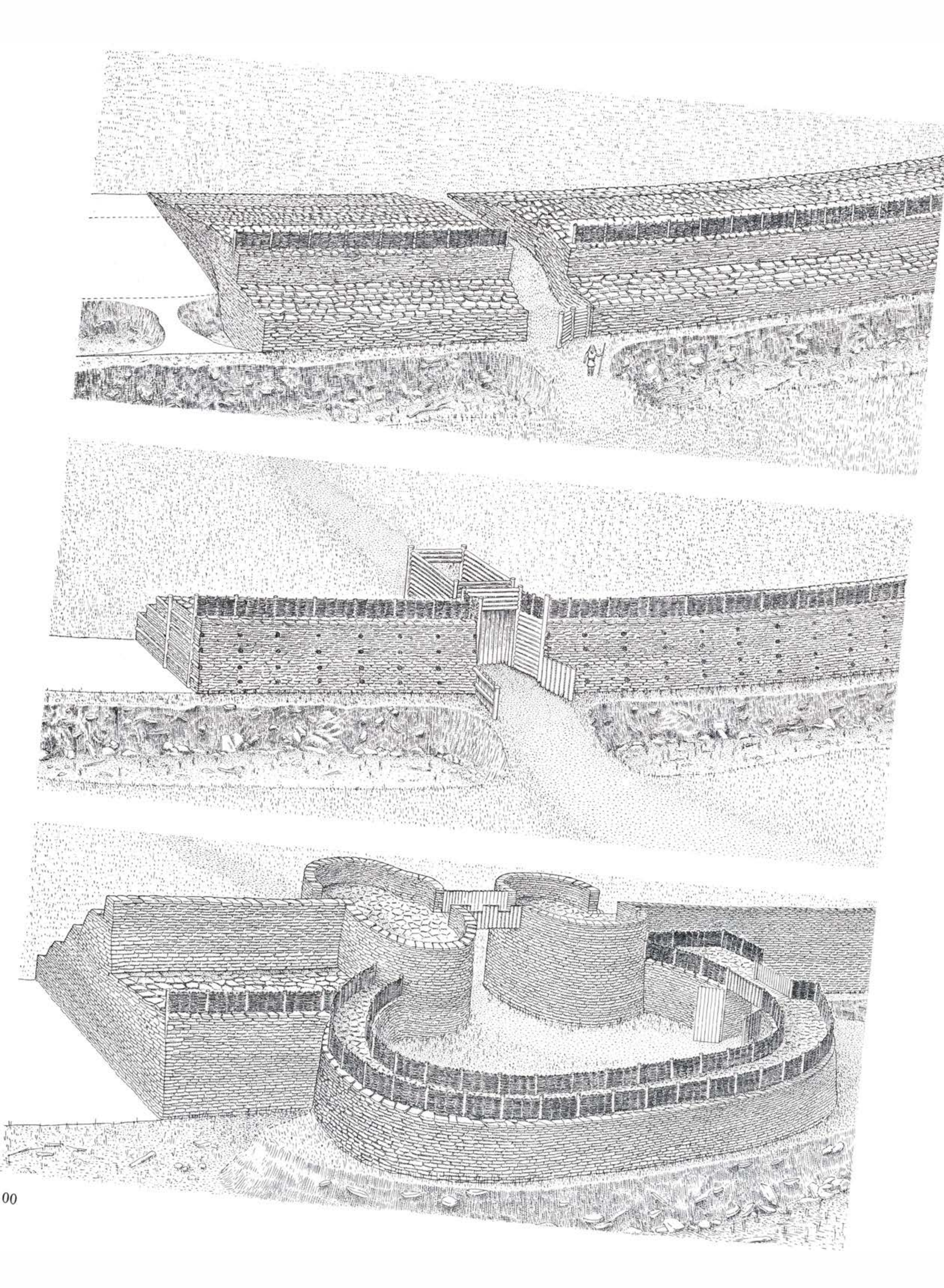
TRES FASES DE OCUPACION están representadas en este plano de Crickley Hill. A la izquierda, cerca del punto más elevado de la colina, se halla el último recinto neolítico, construido alrededor del 2500 a. C. Tenía dos pasadizos de

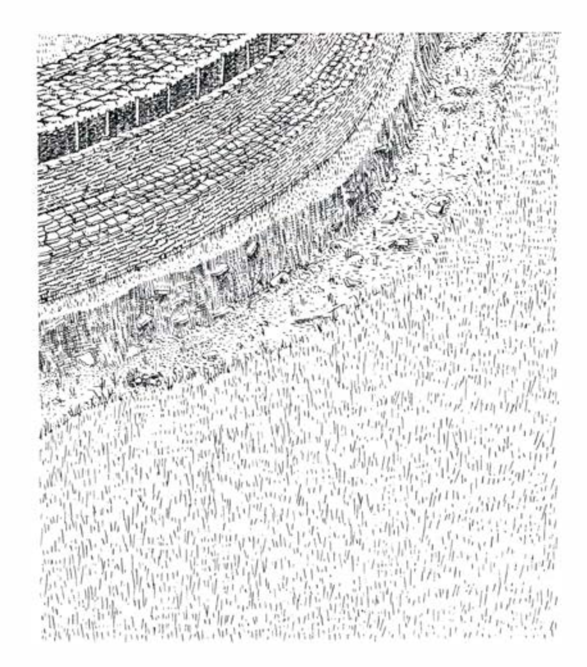
sólo podemos estar seguros de que hubo una importante ocupación neolítica de la cima del cerro y de que sus habitantes hacían útiles de sílex, que por no hallarse en las inmediaciones, han debido importarlo desde un lugar distante, como mínimo a 45 kilómetros. El análisis de secciones finas de cerámica y la microfauna de las fosas y de los agujeros para postes podrán en su momento ayudarnos a distinguir las varias fases.

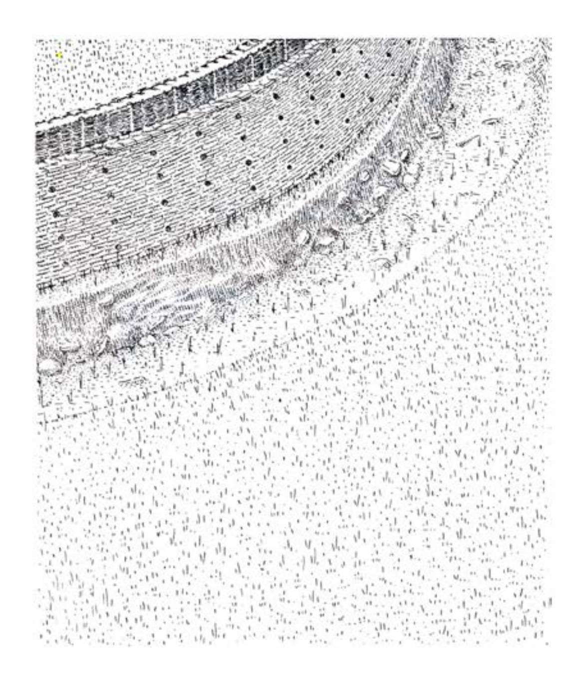
Tras un período de uso, el recinto fue aparentemente abandonado, pero no antes de que los terraplenes fuesen corridos sobre los fosos de donde habían extraído la tierra y quemado sistemáticamente todo el material combustible que se hallaba mezclado con los escombros. Esta medida es indicio de un ritual, pues lo lógico sería que una construcción abandonada se dejase degradar por sí

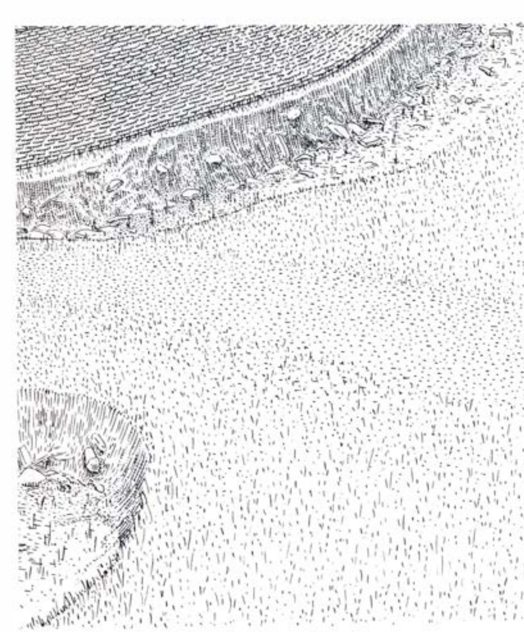


entrada largos y estrechos. Este recinto fue destruido pronto y Crickley Hill quedó abandonada hasta que nuevos colonos construyeron el primero de los dos castros hacia el siglo VI a. C. La nueva fortificación tenía 300 metros de longitud y delimitaba un área de tres hectáreas. La mayor parte del poblado se encontraba a la derecha del "campamento de calzadas" neolítico. El color señala algunos elementos pertenecientes al primer castro, como los hoyos para los postes de sus características casas rectangulares; el negro corresponde a construcciones del segundo castro, realizadas unos años después del fin del primero. La casa redonda mayor, próxima a la entrada, se superpuso en parte a hoyos de postes del centro más antiguo.









misma. Los cortes que realizamos a través de los fosos rellenos indican que en la siguiente fase de ocupación todos los fosos fueron vaciados parcialmente, es de suponer que para reconstruir los terraplenes que los acompañaban. En esta misma fase, se levantaron paredes de piedra seca hasta una altura de un metro a lo largo del borde exterior de los fosos. En la entrada meridional del viejo recinto, las paredes se volvían hacia dentro para flanquear una nueva vía de entrada, ligeramente desviada con respecto a la anterior, pero dando la impresión de restauración y continuidad. También este recinto fue desmantelado más tarde y sus fosos rellenados, pero en algunas zonas los fosos parecen haber vuelto a vaciarse parcialmente, una vez más, antes de que el conjunto entero fuese abandonado.

Esta extraña sucesión de acontecimientos precedió a un fenómeno más extraño aún en la siguiente (y última) fase neolítica. Se excavó un foso considerablemente más ancho inmediatamente por fuera, y paralelo al foso principal interior, de los anteriores fosos con calzadas. A diferencia de los fosos anteriores, el nuevo estaba interrumpido solamente por dos calzadas. En este aspecto se asemejaba al foso defensivo de un castro. Su terraplén tenía también un aspecto militar, pues estaba revestido de un muro vertical de piedra y coronado como mínimo por una empalizada. Sus dos pasadizos de entrada, estrechos y largos, el más pequeño de los cuales medía sólo un metro de anchura, estaban cerrados por puertas firmemente instaladas. Pese a ello, la construcción no era lógica del todo: los constructores decidieron excavar el nuevo foso en sólida caliza antes que vaciar y ampliar los fosos más antiguos. Además, el terraplén del nuevo foso era más bajo de lo que pudiera haber sido, por estar asentado sobre la superficie del antiguo foso rellenado, que no podía por menos de ceder bajo su peso.

¿Fue un hecho deliberado que el nuevo foso, correspondiente a este terraplén, evitase el foso antiguo? Su trazado parece haber sido un manifiesto acto de piedad. Cerca de la entrada meridional los fosos con calzadas tuercen hacia fuera, de un modo irregular, durante un cierto trecho. Los constructores del nuevo foso, que habían empezado excavando una trinchera de medio metro de profundidad aproximadamente, no se percataron de la desviación e irrumpieron en el flanco de uno de los fosos anteriores, que luego volvieron a rellenar. Acabaron su foso hasta la deseada profundidad de dos metros, siguiendo una nueva línea que tenía un profundo quiebro. La superstición de los constructores no sólo se refleja en el trazado del recinto, sino también en unos hoyos pequeños, situados a los extremos del foso, que están cubiertos de grandes losas planas y contienen los huesos de reses sacrificadas.

En esta fase neolítica final, si no en las anteriores, el terraplén y el foso circundaban un poblado. Desde la entrada más pequeña, un camino empedrado flanqueado por una valla, conducía al interior. Al lado del camino encontramos los agujeros de los postes de por lo menos una casa rectangular, y la zona circundante estaba sembrada de restos de ocupación humana. La extensión total disponible para la habitación era de menos de una hectárea, gran parte de la cual queda por excavar. Las variaciones apreciadas en la concentración de los hallazgos arqueológicos hasta el momento presente indican la configuración de un poblado con dos zonas densas, una alrededor de la loma central y otra contigua al interior de los terraplenes. Entre estas dos áreas hay una zona industrial para la manufactura de utensilios de sílex. No conocemos el tamaño de la población, ni siquiera si la estación estuvo ocupada por épocas o de modo permanente. La cuidadosa disposición de sus baluartes le da, sin embargo, un aire de permanencia. Los hoyos de los postes de la casa rectangular fueron rehechos, señal de que la primera casa fue reconstruida, probablemente dos veces, y de que estuvo ocupada por lo menos durante una o dos generaciones.

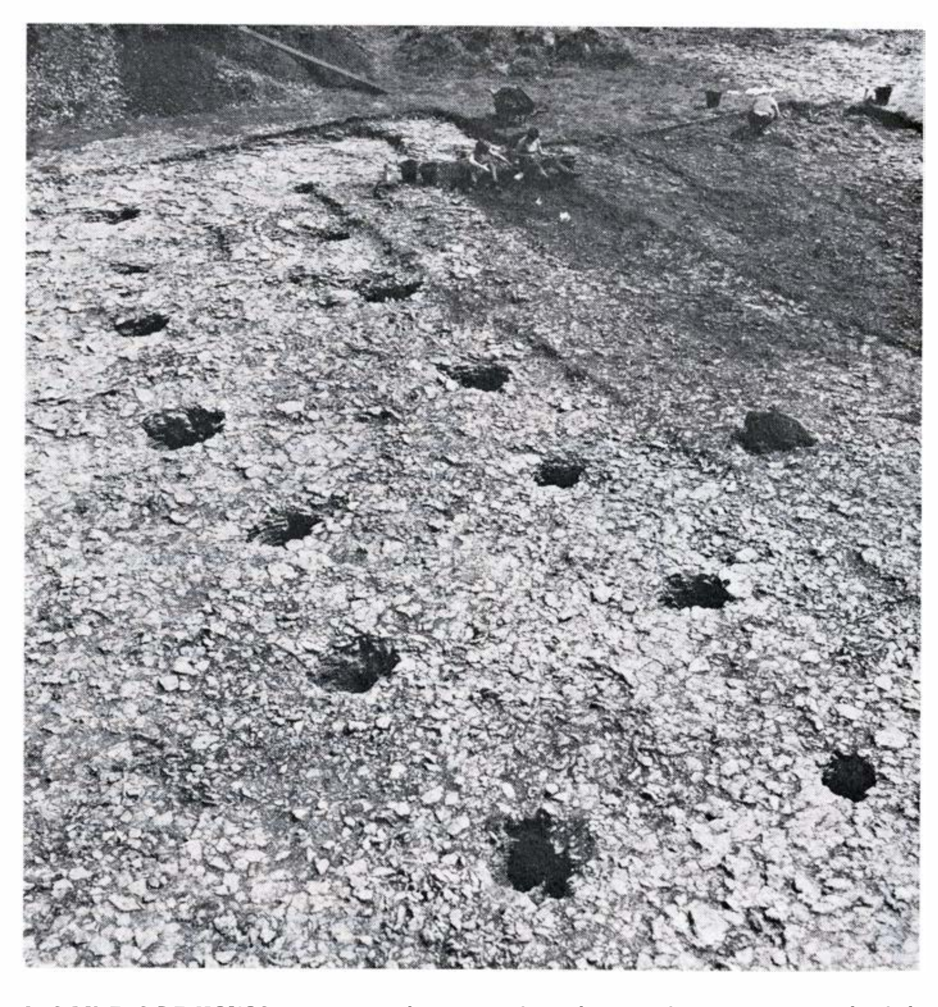
El fin del último poblado neolítico fue repentino y violento. En el foso y en su terraplén, así como en las dos entradas del recinto, encontramos grandes cantidades de puntas de flecha de sílex de tipo foliáceo, algunas de ellas tostadas por el fuego que destruyó las puertas y la casa, y dejó una franja roja a lo largo del dorso del baluarte donde había habido una empalizada de madera. Estas señales inconfundibles de un asedio confirman las observaciones hechas en Carn Brea, en Cornualles, donde se han

ENTRADAS FORTIFICADAS cerraban a tres por lo menos de los poblados prehistóricos del Crickley Hill. Los constructores del último campamento de calzadas neolítico cubrieron los terraplenes del mismo con un revestimiento de piedra e interrumpieron el foso con sólo dos calzadas de acceso, una de las cuales aparece en la reconstrucción de la parte superior. La muralla del primer castro de la Edad de Hierro está reconstruida en el centro. Se componía de muros de piedra seca enlazados por maderos y tenía una sola entrada. Las defensas del segundo castro, reconstruidas abajo, eran más complejas.

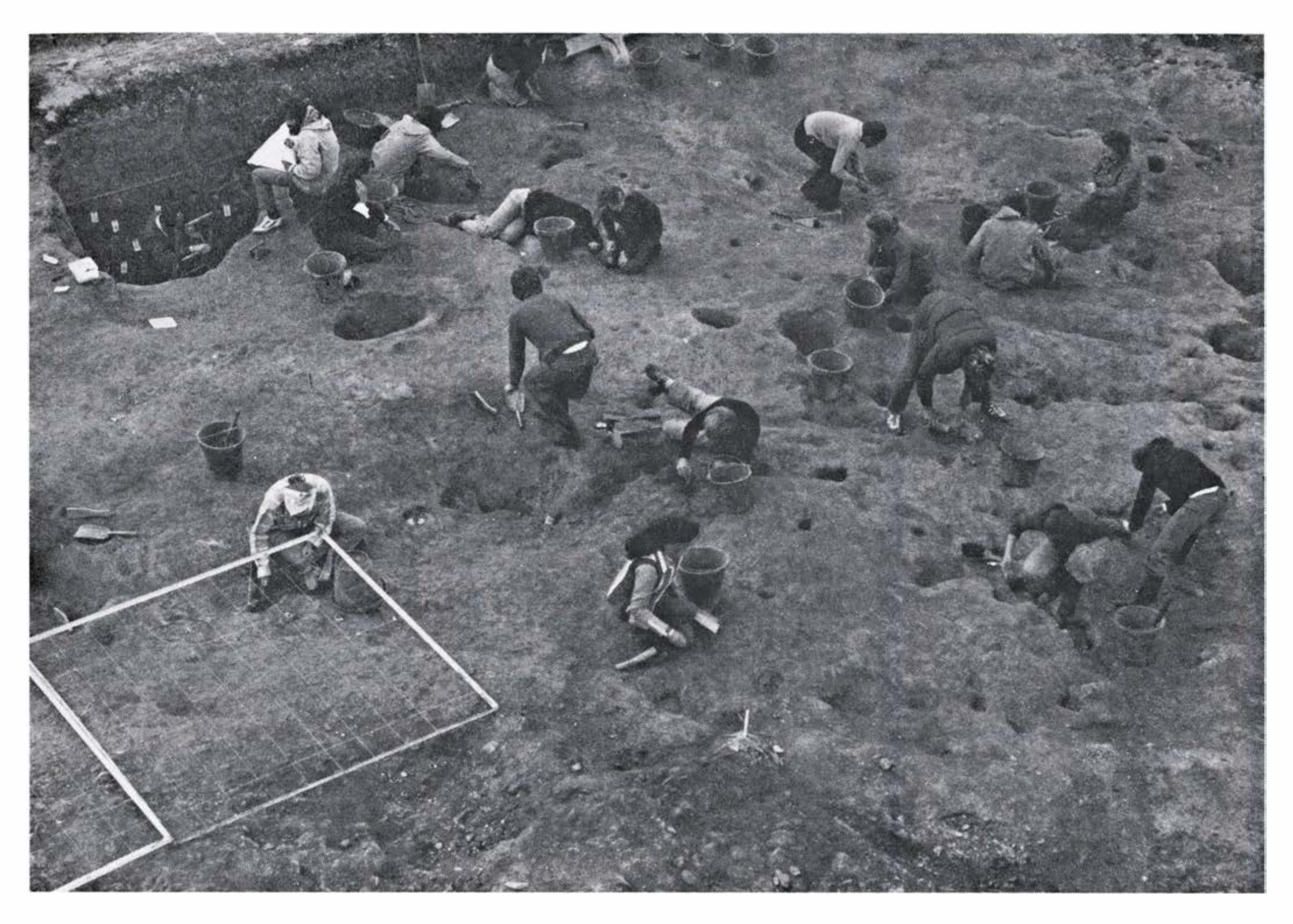
recuperado grandes cantidades de puntas de flecha. Estas puntas de flecha se habían considerado como reliquias de la caza, no de la guerra, porque los arqueólogos han creído durante mucho tiempo que la sociedad del Neolítico inicial en Gran Bretaña era cantonalista y pacífica. El cuadro que ahora está resultando es el de una sociedad sedentaria y estratificada capaz de llevar a término importantes obras comunales; una sociedad que lo mismo construía poblados fortificados que los atacaba. En fecha reciente, las excavaciones de Hambledon Hill, en Dorsetshire, descubrieron en el foso de un modesto poblado neolítico el esqueleto de un hombre con una punta foliácea de flecha en el pecho.

Abandonado tras el incendio, el recinto de Crickley Hill fue cayendo en ruina. La estratigrafía de los depósitos formados a espaldas de la entrada pequeña meridional indican cuál fue el acontecimiento que siguió. Una vez que la pared trasera del muro defensivo se derrumbó, pero antes de que hubiera tiempo para que se acumulase mucho humus en la zona situada inmediatamente detrás, construyeron un gran túmulo alargado, de cuatro metros de ancho y casi cien de largo. El túmulo consistía, casi por completo, en tierra vegetal que había sido laboriosamente arrancada de las laderas de una hondonada natural. Una vez colocada esta tierra en su lugar, los constructores pusieron losas de caliza, una detrás de otra, para formar un bordillo a cada lado del túmulo. El extremo oriental de la construcción, que se superponía a la parte trasera del antiguo terraplén, era semicircular, con el borde señalado por una media luna de piso empedrado. En el centro de la media luna encontramos el hoyo grande de un poste, seguramente la caja para asentar un indicador o un tótem.

Hasta ahora no se han encontrado señales de enterramiento humano en el túmulo, por lo que parece que fue un lugar de ceremonias, quizás el punto al que se dirigía una procesión. No se sabe la época del túmulo, pero ya había adquirido su actual forma erosionada cuando a su lado se construyó una cabaña de la Edad de Hierro. Una vez que el largo túmulo cayó en desuso, la parte



LAS LINEAS DE HOYOS para cimentar los postes indican el tamaño de una casa rectangular de la primera Edad de Hierro. Las paredes laterales se levantaban a dos metros por fuera de los hoyos.



EXCAVADORES VOLUNTARIOS bajo la dirección del autor han excavado casi la mitad de la estación de Crickley Hill en los últimos diez años. Aquí están limpiando la superficie más primitiva neolítica, salpicada de hoyos para postes. Arriba a la izquierda descubren un tramo de un foso.

alta del cerro sólo fue escenario de pastoreo y de caza. La tierra vegetal ha suministrado puntas de flecha de pedernal y fragmentos de cerámica de principios de la Edad de Bronce (posterior al Neolítico pero anterior a la Edad de Hierro), sin ninguna otra señal de ocupación durante casi 2000 años. Después, en el siglo VI a. C. o poco antes, una necesidad ignorada aportó nuevos colonos a la colina y nació el primero de dos castros sucesivos.

La nueva fortificación era mayor que los antiguos recintos, y la mayor parte del poblado se encontraba fuera del área ocupada durante el Neolítico. Detrás de una fuerte muralla con paredes de piedra enlazadas entre sí mediante una compleja armadura de maderos entrelazados, hemos descubierto los grandes hoyos de los postes de seis casas rectangulares, la mayor de ellas de más de 24 metros de longitud. Las casas adoptaban un dispositivo regular a lo largo de un camino que prolongaba la línea de un pasadizo de entrada. Dos casas más del mismo tipo fueron construidas en el último foso neolítico, para entonces casi rellenado, donde los restos del terraplén podían ofrecer cierto abrigo del viento del suroeste, el dominante allí.

Otras casas de la Edad de Hierro, más diseminadas, se hallan en el interior del recinto neolítico. Por lo menos tres tenían un hogar, y las demás pudieron también haber tenido el suyo: la mayor parte de los suelos fueron arrasados durante la ocupación última del castro. Además había, como mínimo, 27 edificios menores, la mayoría de ellos construidos sobre cuatro postes en una disposición cuadrada. Algunos pertenecieron al primer castro y otros a su sucesor; ninguno tuvo hogar y dos estaban llenos de cebada cuando fueron incendiados. Los pozos de almacenamiento son corrientes en los poblados prehistóricos de Gran Bretaña establecidos en terrenos de yeso, pero no ha aparecido ninguno en la caliza de Crickley Hill. Por ello los "cuatro puntales" parecen ser graneros o chozas de almacenamiento. En el primer castro se hallaban diseminados en una extensa zona en torno a los edificios rectangulares principales. Tres edificios de los de cuatro postes, cercanos a la entrada del castro, pudieron haber sido también graneros, pero parece más probable su destino defensivo; un foso pequeño situado junto a las dos cabañas del norte contenía piedras para arrojar con honda.

Cualesquiera que pudieran ser los sistemas defensivos del primer castro, fracasaron claramente a la hora de resistir a sus últimos atacantes, quienes lo incendiaron a conciencia, derribaron sus muros y quemaron la estructura de madera del núcleo del terraplén. Antes de que el castro fuese reconstruido, hubo tiempo suficiente para que algunas capas de restos fuesen arrastradas por la lluvia desde las ruinas, pero no lo bastante para que se formase un suelo erosionado. Como la formación de este suelo requiere más que meses, pero menos que decenios, es evidente que el segundo castro fue construido a poco de la destrucción del primero. Sus constructores modificaron radicalmente las defensas del castro antiguo y su planificación interna. La entrada nueva era de grandes piedras, con bastiones macizos de cara redondeada y una antemuralla con su propia puerta. Los terraplenes, más gruesos que antes, se alzaban como

una impresionante mole. Dentro del castro, una enorme casa redonda, de unos 15 metros de diámetro, dominaba la zona donde se habían alzado en otro tiempo las casas alineadas con regularidad. Como esta casa redonda estaba en línea con la entrada, la calzada empedrada que atravesaba dicha entrada torcía hacia el sur para orlar una fila de casas pequeñas redondas, cada una de ellas de ocho metro de diámetro aproximadamente. Todas las casas redondas tenían un hogar bien construido. Al igual que en el primer castro, los edificios de cuatro postes no tenían dispositivos para fuego, sino que estaban aglomerados a cierta distancia del área de habitación.

A pesar de las refinadas defensas del segundo castro, su suerte no fue mejor que la del primero. Los testimonios indican que fue atacado antes de que hubiese tiempo para que sus nuevas calzadas acusasen las rodadas de los carros y para que sus maderas envejeciesen. El grado de su destrucción está indicado por gruesas capas de carbón a la entrada del castro y por las paredes de los cubos de la muralla enrojecidas por el fuego. Cuando ardió la gran casa redonda, el viento del oeste trazó una franja roja sobre la caliza.

Es improbable que se llegue a conocer algún día la identidad de los atacantes o de los defensores de cualquiera de los dos castros. La cerámica del segundo de ellos recuerda a la de principios de la Edad de Hierro del valle alto del Támesis en Oxfordshire, a unos 30 kilómetros de distancia. La cerámica del primer castro es más difícil de identificar y parece pertenecer a una tradición diferente. Estos útiles, sumados a las pruebas de un cambio evidente y repentino en el diseño de las casas y de las murallas, indican con claridad que las poblaciones de las dos fases eran totalmente distintas una de otra. El más notable de los dos fuertes es el más antiguo. En rigor, sus largas casas rectangulares carecen de paralelos entre los edificios de la Edad de Hierro excavados en Inglaterra, que en los demás sitios han sido casas redondas, y pequeñas construcciones cuadradas o rectangulares de finalidad incierta. En aquella época la casa grande rectangular era corriente en el continente europeo, pero no es necesario suponer que los constructores de Crickley Hill fuesen inmigrantes del otro lado del Canal de la Mancha; un cúmulo de pruebas cada vez mayor atestigua la existencia de casas grandes rectangulares en la Britannia del Neolítico y de la Edad de Bronce. Nuevas excavaciones en otros lugares pueden descubrir más ejemplos que hayan de ser colocados al lado de la casa redonda nativa existente por doquier.

Que nosotros sepamos, Crickley Hill no habría de ser nunca más la sede de un poblado. Las cimas de colinas cercanas fueron fortificadas en el siglo IV a. C. y más tarde, y las aldeas se apiñaron en las tierras bajas a los pies de aquéllas, pero los baluartes de Crickley Hill se abandonaron a su suerte sin la intervención de nadie. A poco más de un kilómetro, en la ladera que mira a Crickley Hill, una mujer del más alto rango fue sepultada con sus joyas poco antes de la conquista romana, pero nada de esa fecha se ha encontrado en nuestras excavaciones. La cima de la colina parece haber vuelto a ser terreno de pastizales, en una soledad que ha continuado hasta la época actual, sólo interrumpida por un metalúrgico que construyó su cobertizo en la entrada del último castro. Suyo pudo haber sido el cerdo que murió aplastado cuando las paredes del bastión se derrumbaron en el siglo II d. C. o más tarde.

Diez años de excavaciones en Crickley Hill han permitido realizar una serie de importantes descubrimientos, incluido el cuadro más claro en Gran Bretaña de un poblado neolítico organizado y fortificado, una curiosa mezcla de elementos rituales y funcionales. No hay razón, sin embargo, para suponer que el yacimiento sea único en la complejidad de su ocupación; ha sido la dimensión de nuestra empresa la que ha dado ocasión a los descubrimientos. De los miles de hectáreas que hay dentro de las murallas de los castros del sur de Inglaterra, son menos del uno por ciento las que se han excavado hasta ahora. A base de una muestra tan pequeña no se pueden hacer deducciones firmes sobre esquemas de poblamiento. Incluso en Crickley Hill, donde en este momento se ha terminado la labor realizada en casi el cuarenta por ciento de la zona interior, apenas empiezan a obtenerse respuestas a importantes preguntas sobre la estructura social, la población y la economía de las varias épocas de ocupación. Disquisiciones coherentes sobre el uso de la tierra y las relaciones de Crickley Hill con los otros lugares habitados en la comarca son prematuras antes de que pueda establecerse la contemporaneidad de Crickley Hill y de las otras estaciones. Queda por hacer mucho trabajo apasionante. Con ayuda de nuestro equipo de trabajadores voluntarios, los veranos de los próximos diez años pueden resultar tan fructíferos como los de los últimos diez años para descifrar las realizaciones de los colonos prehistóricos.

Juegos matemáticos

El número irracional omega parece albergar todos los misterios insondables del Universo

Martin Gardner

Un fascinante artículo inédito, titulado "On Random and Hard-to-Describe Numbers" ("Números aleatorios y números de difícil descripción") ha sido enviado a esta sección por Charles H. Bennett, físico matemático del Thomas J. Watson Research Center, de la International Business Machines Corporation. El artículo comienza recordando ciertas paradojas relativas a números enteros, en particular, la que parece presentarse al decir que un entero bien determinado es aleatorio. La resolución de esta paradoja hace que la aleatoriedad sea propiedad que casi todos los enteros poseen, pero que no puede demostrarse que posean. Este tema fue ya objeto de discusión por Gregory J. Chaitin, también del Thomas J. Watson Research ("*Randomness and Mathematical Proof*", SCIENTIFIC AMERICAN, mayo, 1975). Bennett procede a examinar en qué sentido podría decirse que son aleatorios números como el *pi*, y estudia extensamente el número irracional Ω, hace poco descubierto por Chaitin, que es tan aleatorio que todo plan de apuestas sobre sus dígitos consecutivos daría a largo plazo empate entre ganancias y pérdidas.

El número Ω tiene otras insólitas propiedades. Para empezar, puede definirse con exactitud, pero no es posible computar sus cifras decimales. Y lo más notable, si se conocieran algunos millares de sus primeros dígitos se tendría, al menos en teoría, un procedimiento para contestar casi todas las cuestiones matemáticamente interesantes aún pendientes, en particular, muchas de las proposiciones que, caso de ser falsas, pudieran refutarse mediante razonamientos de número finito de pasos. El análisis que sobre estos temas se expone más abajo ha sido tomado del artículo de Bennett, que se abre con una variante sencilla de la paradoja de Berry. Así llamada en honor a su descubridor, G. G. Berry, bibliotecario en Oxford, y dada a conocer por Bertrand Russell y Alfred North Whitehead en su *Principia Mathematica*, la paradoja es presentada por Bennett en la siguiente forma:

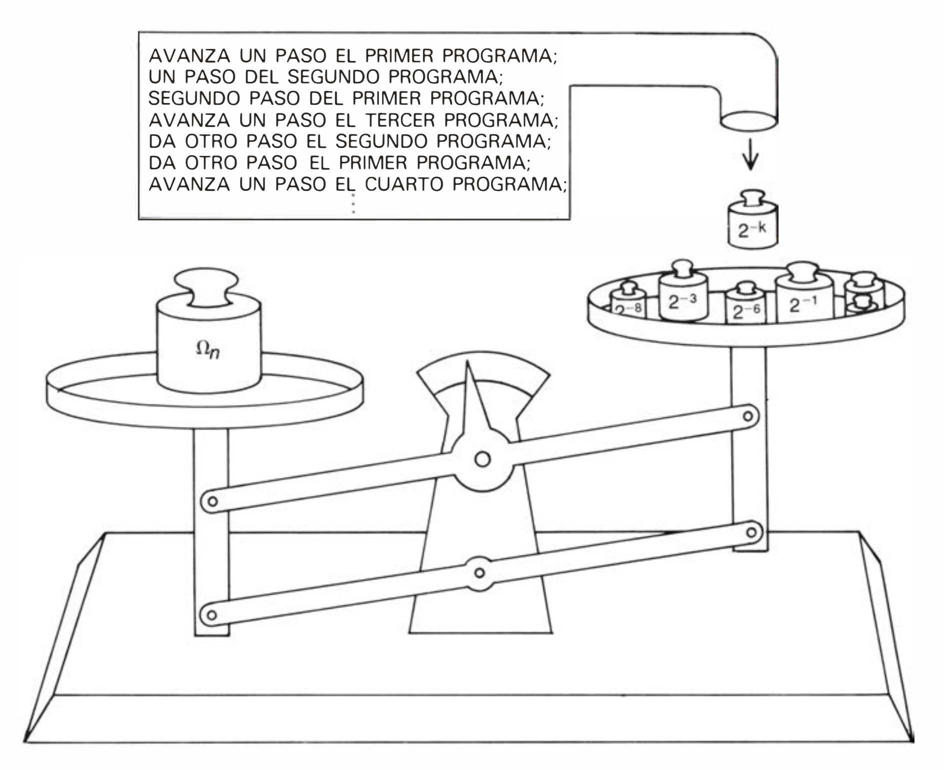
"El número 'treinta y un millones, treinta y un mil treinta y uno' tiene de insólito el ser –o parece ser– el número descrito por la expresión "Primer número imposible de nombrar con menos de once palabras". Ahora bien, esta última frase contiene solamente diez palabras, y por consiguiente, sería contradictorio servirse de ella para nombrar tanto al 31.031.031 como a cualquier otro número. La paradoja de Berry pone de manifiesto que la noción de 'nominabilidad' es intrínsecamente ambigua, y demasiado potente para poder manejarla sin restricciones. En una nota aparecida en *The American Mathematical Monthly* (vol. 52, nº 4, pág. 211, abril, 1945), Edwin F. Beckembach señalaba que, al querer clasificar los números en 'interesantes' y 'corrientes', se tropieza con una paradoja semejante: No pueden existir números 'corrientes', porque de haberlos, el primero de ellos seria 'interesante' precisamente por ser el primero en ser 'corriente'.

"No obstante, la paradoja de Berry puede evitarse, e incluso ser dominada, sin más que restringir el concepto de nominalidad. Un entero se considerará nombrado cuando haya sido computado como salida de un programa de ordenador. Para normalizar el término 'computación' en la definición precedente, se introduce un ordenador simple e ideal llamado 'máquina universal de Turing'. (Puede verse una descripción de las máquinas de Turing en la edición americana de esta sección, SCIENTIFIC AMERICAN, junio, 1971). Esta máquina es capaz de admitir un programa dado por una sucesión de ceros y unos grabados en una cinta de entrada, y escribir el resultado de sus cálculos –también en forma de sucesión de cifras binarias– sobre una cinta de salida, al término de su computación. Durante el proceso de cómputo se utiliza una tercera cinta para almacenar resultados intermedios. (El utilizar cintas distintas para entrada, salida y memoria se debe más a comodidad conceptual que a necesidad; en las primeras máquinas de Turing una misma cinta servía para estos tres propósitos.) Como ya es sabido, la máquina universal de Turing puede realizar toda tarea factible en el más perfecto ordenador digital, si bien mucho más lentamente. Con mayor generalidad, la máquina de Turing puede realizar incluso la más complicada manipulación de informaciones numéricas o simbólicas, siempre que tal manipulación venga expresada como sucesión finita de pasos elementales, donde cada paso siga al precedente por un proceso puramente mecánico, sin intervención de raciocinio ni de azar.

"Un entero x puede ser nombrado ahora especificando una sucesión binaria p que, utilizada como programa de entrada en una máquina de Turing, haga que la máquina calcule x, entregándolo como única salida, y se detenga luego. No cabe duda de que el programa p describe verdaderamente a x; por tanto, la máquina universal de Turing proporciona un lenguaje flexible y exento de ambigüedades, mediante el cual es posible describir números por cualquiera de los procedimientos que permitan calcularlo efectivamente. (Por ejemplo, el número 523 podría ser descrito como 99-ésimo término de la serie de número primos, como $(13 \times 41) - 10$, o de un modo más directo, mediante la sucesión binaria 1000001011.) Mediante este lenguaje es posible nombrar a todo entero, pues incluso los enteros que no posean propiedades que los distingan podrían siempre describirse sin más que dar su expresión binaria.

"Volvamos ahora a la clasificación de los números en interesantes y corrientes. Podemos definir –esta vez sin paradojas– como interesantes aquellos números computables mediante programas que contengan muchos menos bits (dígitos binarios) que el propio número. La posibilidad de una tal descripción abreviada sería reflejo de alguna característica peculiar, que distinguiría al número. Así pues, con esta definición, $2^{65.536} + 1$, el primer millón de cifras decimales de pi y $(17!)!$ serían números interesantes. (El signo de exclamación es el signo de 'factorial'; $n!$ es igual a $1 \times 2 \times 3 \times \ldots \times n$.) Recíprocamente, los números 'corrientes', o números aleatorios, serían aquellos que no pudieran comprimirse de un modo significativo, es decir, cuya descripción más breve tuviera sensiblemente el mismo número de bits que el propio número. Esta definición algorítmica del carácter aleatorio,

expresado como incomprensibilidad, que se menciona en el artículo de Chaitin antes citado, fue puesta a punto por diversos matemáticos en la década de 1960, entre ellos, Ray J. Solomonoff y Chaitin, de los Estados Unidos, y A. N. Kolmogorov, de la Unión Soviética. Al igual que la mayoría de los números, serían aleatorios en el sentido intuitivo del término; también la mayoría de los números son incompresibles –o casi– pues hay demasiados pocos programas cortos de qué echar mano. Dicho de otra forma, incluso sin desperdiciar programas (como sucedería al computar un mismo resultado mediante varios programas), sólo una pequeña parte de los enteros *n* bits podría describirse mediante programas más breves –ni aún en unos cuantos bits– que los propios números.



Forma de usar los primeros **n** *bits de* Ω *para resolver el problema de detención de todos los programas hasta* **n** *bits*

"Utilizando esta definición de entero aleatorio, Chaitin puso al descubierto un hecho sorprendente: si bien casi todos los enteros son aleatorios, dentro de un sistema axiomático coherente dado tan sólo se podrá probar que un número finito de ellos lo son. Este resultado, que es una de la formas del famoso teorema de incompletitud de Gödel, implica en particular que en un sistema cuyos axiomas y reglas de inferencia puedan quedar descritos mediante *n* bits no se podrá demostrar el carácter aleatorio de números de longitud mucho mayor que *n* bits. La demostración de este aserto dada por Chaitin se apoya en una versión computerizada de la paradoja de Berry: Supongamos que en un sistema de demostración descriptible mediante un pequeño número de bits pudiera establecerse el carácter aleatorio de algún entero de gran número de bits. Se podría entonces diseñar un pequeño programa de máquina de Turing basado en tal método de demostración, programa que produciría como salida el entero grande. Sin embargo, si el entero grande fuese verdaderamente aleatorio, no podría ser salida de ningún programa pequeño, con lo que se llega a una contradicción.

"Con mayor precisión, el programa de la máquina de Turing llevaría incorporada una rutina de supervisión, es decir, un subprograma, cuya longitud supondremos sean *c* bits. Esta rutina utilizaría los *n* bits destinados a axiomas y reglas de inferencia para ir generando sistemáticamente todas las posibles demostraciones que pudieran deducirse de los axiomas, según el número creciente de pasos deductivos. Generaría primero todas las demostraciones de un solo paso, después, todas las de dos, y así sucesivamente. Tras ser generada cada demostración, la rutina comprobaría si tal demostración establece la aleatoriedad de algún entero de longitud considerablemente mayor que $n + c$ bits. De encontrar alguna demostración así, la rutina de supervisión ordenaría imprimir el gran entero determinado por la demostración, deteniendo a continuación todo el proceso de cómputo. Empero, la longitud total del programa de la máquina de Turing sería $n + c$ bits (correspondientes a los axiomas, las reglas de inferencia y la rutina de supervisión). Dicho de otra forma, un programa de longitud $n + c$ bits produciría como salida un entero bien determinado, que por la definición algorítmica de aleatoriedad no puede ser generado por ningún programa de máquina formado por tan sólo $n + c$ bits. La única vía de escape a esta contradicción está en concluir que, o bien el sistema axiomático es inconsistente (es decir, que dentro de él pueden demostrarse enunciados no verdaderos), o que el proceso de generación sistemática de demostraciones ha de proseguir indefinidamente, sin desvelar nunca una demostración de aleatoriedad para ningún entero de tamaño mucho mayor que $n + c$ bits. La paradoja de Berry, en su primitiva versión parecía ser un engorro, pues ponía en duda la noción, a primera vista significativa, de entero aleatorio. En su versión computerizada, la paradoja de Berry contornea esta noción del necesario margen de indemostrabilidad, permitiendo así definir el concepto sin contradicciones.

"Mucho antes de que el concepto de número entero aleatorio fuese siquiera tomado seriamente en consideración, Emile Borel, Richard von Mises, y otros matemáticos buscaron definir y encontrar ejemplos de números reales aleatorios, o lo que es equivalente, sucesiones aleatorias infinitas de dígitos, sean binarios o decimales. Se ha conjeturado que ciertos números irracionales, tales como pi, *e*, $\sqrt{2}$, que se presentan de forma natural en matemática, son aleatorios, en el sentido de que cada una de sus cifras, y más aún, cada tramo de cifras de longitud fija, se presenta con la misma frecuencia en su desarrollo decimal. Las sucesiones con esta propiedad se llaman normales. No es difícil demostrar que ningún número racional es normal, cualquiera que sea la base de numeración en que demos su desarrollo, y también, que casi todos los números irracionales han de ser normales en todas las bases. No obstante, hasta el momento no ha podido demostrarse que ninguno de los números irracionales clásicos sea normal, aunque la evidencia estadística existente apoya en términos generales la hipótesis de que sí lo son.

"Por otra parte, es fácil construir números irracionales 'artificiales' cuyo carácter normal es demostrable, a pesar de que sus cifras siguen pautas triviales y

diáfanas. El más famoso de estos números fue inventado por D. G. Champernowne a comienzos de la década de 1930: 0,12345678910 11 12 13 14 15 16 17 18 19 20 21 22 23 24... Para este número ridículamente sencillo, formado por la sucesión creciente de los números naturales en base 10, se ha podido probar no sólo que es irracional y normal (en base 10) sino también, que es trascendente. (Los números trascendentes *no* son raíces de ecuaciones algebraicas de coeficientes enteros.) En los primeros tramos del número de Champernowne se producen importantes desviaciones respecto del carácter normal, pero las diferencias entre las frecuencias tienden a cero conforme aumenta el número de dígitos examinados. Según parece, se ignora si este número continúa siendo normal en otras bases de numeración.

"Si bien la sucesión de dígitos del número pi puede ser aleatoria, en el sentido de que puede ser normal, tal sucesión no es impredictible. Dicho con otras palabras, un jugador suficientemente hábil que fuese apostando sobre las sucesivas cifras de pi podría eventualmente inferir la regla de formación, y a partir de ahí ganar todas las apuestas. Lo mismo vale para el número de Champernowne. ¿Existirá alguna sucesión tan aleatoria que ninguna estrategia de apuestas computable, que vaya apostando sobre las sucesivas cifras consiga desequilibrar a largo plazo pérdidas y ganancias? Todo número normal en este sentido fuerte habrá necesariamente de ser normal en todas las bases de numeración. Uno de los resultados fundamentales de la teoría de probabilidad establece que, de hecho, casi todos los números reales son aleatorios en este sentido; pero no es fácil encontrar ejemplos concretos de números así. Además, hay un sentido en el cual ningún número real específicamente definible puede ser aleatorio, dado que hay una infinidad no numerable de números reales (es decir, el conjunto de los números reales es demasiado grande para poder emparejarlo biunívocamente con los enteros positivos), mientras que solamente hay una infinidad numerable de posibles definiciones. Con otras palabras, el mero hecho de ser definible le confiere al número real definido un carácter atípico. En este caso, sin embargo, el problema consiste nada más en determinar un número cuya atipicidad no pueda demostrarse por métodos computacionales, o sea, constructivos. En particular, un número así no podrá ser computable a partir de su definición, pues si lo fuera se podría diseñar una estrategia de apuestas que fuese perfecta.

"El número irracional Ω, ideado por Chaitin, tiene, entre otras notables propiedades, la de ser aleatorio en este sentido fuerte. Sin embargo, para comprender por qué, es necesario examinar brevemente un problema no resoluble, clásico en teoría de computabilidad, conocido como problema de terminación, es decir, el problema de distinguir entre los programas de ordenador que llegarían espontáneamente a detener la máquina y aquéllos que la harían funcionar indefinidamente. Aparte gruesos errores de programación, que provocarían la detención o no detención del programa por razones triviales, los programas terminan cuando han conseguido realizar lo que se proponían, por ejemplo, cuando han llegado a calcular el 99-ésimo número primo, o el primer millón de cifras decimales de pi. Recíproca-

mente, el programa continuará indefinidamente cuando la tarea encomendada no tenga término, como por ejemplo, calcular todos los números primos o determinar un mapa planar que no pueda colorearse con sólo cuatro colores, con la única condición de que ningún par de regiones fronterizas entre sí sean del mismo color.

"A primera vista, el problema de detención o terminación podría parecer resoluble. Después de todo, siempre sería factible probar que un programa llega a detenerse, sin más que hacerlo funcionar suficiente tiempo. Además, hay muchos programas donde es fácil establecer si llegarán o no a término, sin necesidad de ensayarlos. (Por ejemplo, el famoso teorema de los cuatro colores, que establece que para la tarea de iluminar un mapa como antes se dijo nunca serán necesarios cinco colores, se demostró por fin en 1976, y tal demostración garantiza que el programa de coloreado de mapas nunca concluirá.) La dificultad no reside, pues, es resolver ciertos casos particulares del problema de detención, sino en resolverlo en general. A. M. Turing, matemático inglés inventor de la máquina de su nombre, demostró que no existe una receta general que permita decidir cuánto tiempo ha de estar en funcionamiento un programa para que llegue a revelar si la detención se producirá o no. También demostró Turing que no existe ningún sistema coherente de axiomas lo bastante fuerte como para decidir en todos los programas si llegarán o no llegarán a término, sin necesidad de pasarlos. El carácter insoluble del problema de terminación puede deducirse (de hecho, es equivalente a él) del carácter aleatorio, en el sentido de incomprensibilidad de Chaitin y Kolmogorov, que casi todos los enteros poseen, aunque no se puede demostrar que posean.

"Supongamos ahora que a la máquina de Turing, en lugar de un programa bien definido, se le introdujera una sucesión aleatoria de bits, lo que podría conseguirse lanzando al aire una moneda cada vez que la máquina solicitase un bit de su cinta de entrada, dándole un 1 o un 0 según que la moneda haya salido cara o cruz. Actuando así podemos suscitar una curiosa cuestión: al comenzar este proceso, ¿cual es la probabilidad de que la máquina llegue a detenerse?

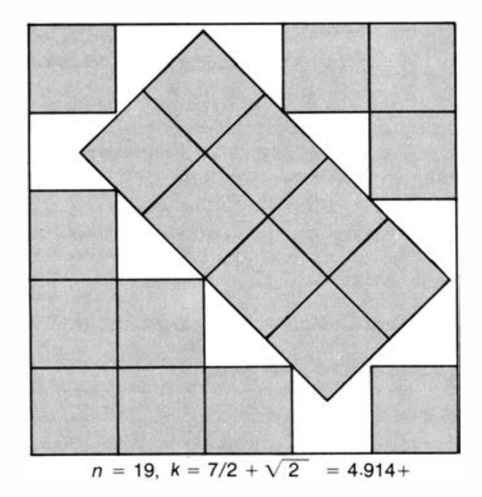
"La solución es el número Ω de Chaitin. Dado que el valor de Ω depende de la máquina universal de Turing que se esté utilizando, Ω no es una única constante universal, como lo es pi. Sin embargo, para una máquina dada, Ω es un número irracional bien definido, comprendido entre 0 y 1, y cuya interpretación natural es la probabilidad de que la máquina llegue a detenerse al ser alimentada con un programa aleatorio. Es muy verosímil que un programa tomado al azar ordene al computador realizar algo imposible o absurdo, así que o bien la máquina se detendrá inmediatamente al tropezar con un error, o entrará en círculos viciosos, obedeciendo una corta serie de instrucciones. En casi todos los ordenadores predomina la primera conducta, y así, por ser la probabilidad de detención cercana a 1, el desarrollo decimal de Ω comienza por varios nueves consecutivos. No obstante, puede demostrarse que pronto la sucesión de cifras de Ω deja de ajustarse a pauta ninguna, y que, al cabo, supera a toda estrategia de colocación de apuestas

que sea computable. Además, es aleatorio en el sentido de Chaitin y Kolmogorov, o sea, es incomprensible.

"Empero, la propiedad de ser aleatorio no es la más notable que Ω posee. Después de todo, comparte esa propiedad con la inmensa mayoría de los números reales. La verdad es que si se conocieran unos cuantos millares de las primeras cifras decimales de Ω, bastarían, al menos en teoría, para decidir muchas de las cuestiones hoy pendientes en matemática. Tal propiedad, lo mismo que la invulnerabilidad del número Ω frente a estrategias de apostado, se debe al compacto método con que Ω codifica soluciones del problema de terminación.

"Probablemente, el más famoso de los problemas hoy pendientes sea el 'último teorema de Fermat', donde se afirma que la ecuación $x^n + y^n = z^n$ carece de soluciones cuando n es mayor que 2. Pierre de Fermat anotó este aserto en el margen de un libro de teoría de números, añadiendo haber descubierto una notable demostración, pero que el margen era demasiado estrecho para darle cabida. Fermat murió sin publicar su demostración, y tres siglos de esfuerzos de otros matemáticos no han conseguido producir ni demostración ni refutación.

"El último teorema de Fermat, al igual que muchas famosas conjeturas matemáticas aún no resueltas, es una afirmación de inexistencia, y por tanto, para refutarla bastaría encontrar un solo contraejemplo finito, a saber, un conjunto de cuatro enteros x, y, z y n que resolvieran la ecuación. Estas conjeturas refutables por procedimientos finitos son equivalentes al aserto de que cierto programa de ordenador, cuya tarea sea la búsqueda sistemática del objeto presuntamente inexistente, nunca llegará a detenerse. Otra famosa proposición refutable en términos finitos es la conjetura de Goldbach, que afirma que todo número par es suma de dos números primos.



Solución al problema de empaquetado, del mes pasado

"Otro tipo de conjetura finitamente refutable que ha ocupado muy importante papel en la historia de la matemática afirma que una determinada proposición es independiente de un sistema axiomático dado, o sea, que la proposición no puede ser demostrada ni refutada. Las más famosas proposiciones de este tipo son el postulado de las paralelas, que afirma que, en el plano, por un punto dado puede trazarse una y exactamente una paralela a una recta dada, y la hipótesis del continuo, que afirma que no existe ningún número transfinito comprendido entre aleph-sub-cero (número de elementos del conjunto de los números naturales) y aleph-sub-uno (número de elementos del conjunto de números reales). Durante el siglo pasado se pudo demostrar que el postulado de las paralelas era independiente de los restantes axiomas de la geometría euclídea, y en este siglo se consiguió demostrar que la hipótesis del continuo era independiente de los axiomas de la teoría de conjuntos. La independencia de una proposición P respecto de un sistema axiomático dado es equivalente a la no terminación de un programa que vaya generando sistemáticamente demostraciones a partir de los axiomas, buscando demostración o refutación de P.

"No todas las conjeturas famosas son finitamente refutables. Por ejemplo, ninguna cantidad finita de pruebas favorables, por extensa que sea, permitirá decidir si el número pi es normal, o si hay infinitos números primos gemelos (números primos impares consecutivos, como 11 y 13 u 857 y 859) o si la conjetura $P \neq NP$ de la teoría de complejidad es verdadera. [La conjetura $P \neq NP$ afirma que existen problemas matemáticos donde la validez de soluciones halladas 'por intuición' puede verificarse rápidamente, aunque para los cuales no existan procedimientos algorítmicos rápidos.] Tales conjeturas no son equivalentes a problemas de detención, pero existen buenas razones para creer que muchos de ellos podrían resolverse indirectamente, demostrando conjeturas más fuertes que sean finitamente refutables. Por ejemplo, se conocen hoy muchos números primos gemelos, y la prueba empírica hace pensar que la separación entre ellos crece con bastante lentitud. Por consiguiente, la conjetura sobre números primos gemelos puede considerarse como forma innecesariamente débil de una proposición más fuerte y quizá cierta, concerniente a la separación entre pares consecutivos de números primos gemelos, que pudiéramos enunciar así: entre 10^n y 10^{n+1} existe al menos un par de enteros primos gemelos. Esta versión reforzada es equivalente a la no detención de un programa que buscase lapsos excesivamente largos (de longitud mayor que un factor 10) en la distribución de números primos gemelos. (Es importante observar que ciertas cuestiones matemáticas no son reducibles a problemas de detención, como ocurre con ciertas cuestiones relativas al propio número Ω. No obstante, este tipo de cuestiones irreducibles propenden a ser bastantes artificiosas y auto-alusivas.)

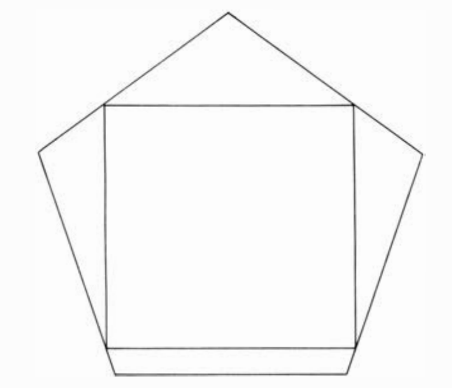
"Las conjeturas interesantes, lo mismo que los números interesantes, acostumbran a tener descripción concisa. Cuesta trabajo imaginar una conjetura interesante desde el punto de vista matemático, y finitamente refutable, de tan verbosa descripción que no pueda codificarse como problema de detención de un programa breve, de unos pocos miles de bits de longitud. Por consiguiente, las respuestas a todas las conjeturas interesantes de este tipo, incluidas las no formuladas todavía, serían en principio accesibles si se dispusiera de alguna especie de 'oráculo' capaz de resolver el problema de detención para todos los programas de hasta unos cuantos millares de bits. El número de programas a considerar todavía sería enorme; por ejemplo, hay aproximadamente 2^{1000} programas de longitud menor que 1000 bits. Parece, pues, que un oráculo capaz de resolver correctamente tantos problemas tendría que ser, o bien muy listo, o bien disponer de una enorme cantidad de información almacenada. De hecho, por la muy densa forma en que Ω codifica soluciones del problema de detención, unos cuantos de sus primeros millares de bits servirían de oráculo perfecto.

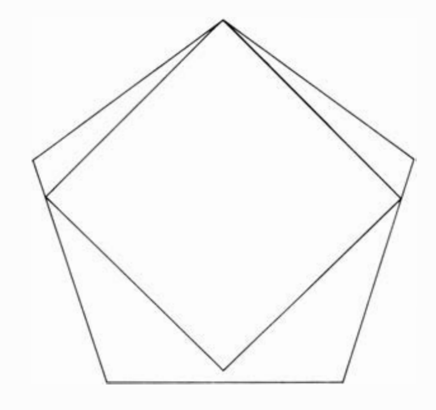
"¿De qué forma pueden recobrarse a partir de Ω soluciones de problemas concretos de terminación de programas? Dado que Ω está definido como la probabilidad global de detención de un ordenador frente a un programa de entrada aleatorio, Ω puede ser considerado como suma de las probabilidades individuales de todos los cómputos que llegan a detenerse. Cada programa que produzca un cómputo finito aporta a la suma la probabilidad de resultar elegido (como si hubiese salido por casualidad) al irse obteniendo los bits del programa de entrada por lanzamiento de una moneda. La probabilidad de generar un programa de k bits de longitud al lanzar k veces la moneda es $1/2^k$. Por consi-

guiente, si al alimentar la máquina de Turing, bit por bit, con este programa se consigue que la máquina se embarque en un cómputo que provoque su detención tras reclamar y leer exactamente k bits del programa, entonces la aportación a Ω efectuada por este programa es exactamente $1/2^k$. (Se considera que los programas que exigen al ordenador leer más o menos bits que los que el programa contiene no determinan la detención de la máquina. Se hace así para contar una sola vez la aportación a Ω de cada programa.)

"La ilustración de la página 107 muestra de qué forma pueden utilizarse los n primeros bits de Ω en la resolución del problema de detención para todos los programas de longitud no mayor que n bits. Como Ω es un número irracional, el número Ω_n fomado por sus primeros n bits da una ligera subestimación del verdadero valor: Ω_n es menor que Ω y éste, menor que $\Omega_n + 1/2^n$, a su vez. Al objeto, de resolver el problema de detención para todos los programas de n bits se comienza una búsqueda sistemática, aunque interminable, de todos los programas que producen detención, de todas las longitudes, haciendo funcionar primero un programa y luego otro, y otro, durante tiempos más y más largos, hasta que se hayan descubierto suficientes programas que lleguen a producir detención como para que la probabilidad acumulada de detención supere a Ω_n.

"Un método para visualizar este proceso consiste en imaginar una balanza en cuyo platillo izquierdo se coloca un peso igual a Ω_n. Como muestra la ilustración, los programas se van pasando por un procedimiento reiterativo: primero se da entrada al primer paso del primer programa, después un paso del segundo programa y otro paso del primero; a continuación, un paso del tercer programa, otro paso del segundo y otro del primero, y así sucesivamente. Cada vez que un programa de longitud k detiene la máquina, se echa un peso igual a $1/2^k$ en el platillo de la derecha de la balanza, porque la probabilidad de que, al alimentar el ordenador de acuerdo con los resultados de consecutivos lanzamientos de la moneda, resulte ejecutado un tal programa es $1/2^k$. Prosiguiendo suficientemente, la balanza terminará por inclinarse hacia la derecha, pues el peso total de los programas que producen detención –la probabilidad de detención es Ω– es un número irracional comprendido entre Ω_n y $\Omega_n + 1/2^n$. Para entonces se habrán descubierto ya muchísimos programas que determinan detención, algunos de más de n bits, y otros más cortos; muchos programas que de continuar funcionando llegarían a detener el cómputo todavía no lo han hecho. No obstante, una vez desequilibrada la balanza hacia la derecha ya no quedarán programas de longitud n o menor, capaces de detener el ordenador, pues de existir alguno más, el valor de Ω superaría a su cota superior ya establecida, $\Omega_n + 1/2^n$. O sea, la detención producida por un nuevo programa de n bits o menor implicaría modificar uno de los dígitos de Ω ya conocidos.





Solución del problema del pentágono, del mes pasado

"Si este pantagruélico cómputo fuese realizado hasta alcanzar una estimación de Ω suficientemente exacta, digamos de unos 5000 bits o más, resultaría que entre todos los programas cuya suerte estaría echada se encontraría uno cuya no detención significaría haber comprobado el teorema de Fermat. Se encontrarían también otros programas que zanjarían la paradoja de Goldbach, así como todas las demás conjeturas de enunciado breve y factibles de refutar en número finito de pasos. Entre ellos se contarían también programas cuya no detención zanjaría, a buen seguro, muchas conjeturas no finitamente refutables, como las relativas a la normalidad del número pi, la existencia de números primos gemelos, y la cuestión de si $P \neq NP$, probando en su lugar proposiciones más fuertes, finitamente refutables.

"Refiriéndonos nuevamente a los sentidos en que el propio Ω es aleatorio –su carácter incomprensible y su invulnerabilidad frente a estrategias para apostar sobre sus cifras– puede parecer extraño que Ω contenga tanta información sobre el problema de terminación y a pesar de ello sea computacionalmente indistinguible de cualquier sucesión aleatoria desprovista de significado. En realidad, Ω constituye un mensaje saturado de información, que parece ser aleatorio porque ha sido expurgado de toda redundancia; un mensaje constituido por información que no puede ser obtenida por ningún otro procedimiento.

"Para poner de relieve esta absoluta carencia de redundancia que Ω posee, examinemos un método más tradicional de codificar el problema de detención mediante un número irracional no computable. Definimos K como el número real cuyo n-ésimo bit es 1 o 0, según que el n-ésimo programa llegue o no a detención. De hecho, suele decirse que K es un oráculo para el problema de detención; pero se trata de un oráculo cuya información está muy diluida, en el sentido de que sus 2^n primeros bits contienen sensiblemente la misma información que los primeros n de Ω, a saber, la suficiente para resolver el problema de detención para programas de n bits o menores. Además, K no es invulnerable frente a estrategias de apostado, justamente a causa de su dilución. Por ejemplo, en una considerable porción de todos los programas puede demostrarse fácilmente si se produciría detención o no, y ello por razones triviales. Los correspondientes bits de K son entonces predecibles, y un jugador que apostase sólo en esos casos, 'pasando' en los demás, podría ganar regularmente. Además, ni siquiera los bits no predecibles de K lo son totalmente. Con frecuencia puede observarse que dos programas distintos están en realidad analizando un mismo problema desde distintos ángulos. Si uno de ellos llegara a detenerse sería segura la detención del otro, y un jugador que hubiera 'pasado' en el primero, no sabiendo entonces cuál debería ser su apuesta, podría más tarde apostar sobre seguro en el segundo.

"La propiedad de ser Ω incomprensible e inmune frente a estrategias de apostado deriva de su compacta codificación del problema de detención. Dado que los primeros n bits de Ω resuelven el problema de detención para todo programa no mayor que n bits, constituyen un 'axioma' suficiente para demostrar la incomprensibilidad de todos los enteros incomprensibles expresables mediante n bits o menos. Si fuese posible computar

Ω_n mediante un programa significativamente menor de n bits, resultaría que un programa de tamaño análogo bastaría para descubrir y hace imprimir el primer entero incomprensible de n bits, lo que es contradictorio. O sea, como Ω_n proporciona información suficiente para calcular un entero incomprensible y bien determinado de n bits, resulta que el propio Ω_n ha de ser incomprensible.

"A través de la historia, místicos y filósofos han buscado una clave que condensara la sabiduría universal, un texto, una fórmula finita que proveyera de respuesta a toda cuestión. La Biblia, el Corán, I Ching, ha sido utilizados para profecía y predicción, y junto con los libros secretos de Hermes Trismegisto y la Cábala judía medieval, son ejemplo de esta creencia y manifestación de esta esperanza. Tales fuentes de sabiduría han estado tradicionalmente protegidas contra su aplicación directa e indebida por la dificultad en encontrarlos, la dificultad de comprenderlos y la peligrosidad de aplicarlos. Además, tienden a responder más –y más profundas– cuestiones de las que desea preguntar quien recurre a ellos. El libro esotérico es, lo mismo que Dios, simple, pero indescriptible. Es omnisciente, y transforma a todo aquel que lo conoce. Hoy se considera superstición el recurso a estos textos clásicos para predecir sucesos mundanos, y no obstante, aunque en diferente sentido, la ciencia está buscando su propia Cábala, un sistema conciso de leyes naturales capaces de explicar todos los fenómenos. En matemática, donde ya no hay esperanza de que ningún sistema de axiomas permita demostrar todas las proposiciones verdaderas, la meta podría ser una axiomatización concisa de todos los enunciados verdaderos 'interesantes'.

"En muchos sentidos, Ω es un número cabalístico. Se puede tener conocimiento de él a través de la razón humana, pero no puede ser conocido. Para conocerlo con detalle se tendría que aceptar la sucesión no computable de sus cifras por profesión de fe, como las palabras de un texto sagrado. El número encierra enorme cantidad de sabiduría en muy poco espacio, en tanto que unos cuantos millares de sus cifras, que podrían escribirse en una hoja de papel, contienen las respuestas de más cuestiones matemáticas de las que podrían escribirse en el universo entero; entre ellas, todas las conjeturas finitamente refutables que son de interés. La sabiduría contenida en Ω carece de utilidad precisamente por ser universal: el único método conocido para extraer la solución de uno de los problemas de detención, por ejemplo, la conjetura de Fermat, a partir de Ω sería embarcarse en una ingente tarea de computación que al mismo tiempo proporcionaría la solución de todos los demás problemas que admitan enunciados breves; tarea con mucho excesivamente grande para acometerla. Por pura ironía, sin embargo, si bien es imposible computar Ω, podría resultar accidentalmente generado por un proceso aleatorio, como una serie de lanzamientos de monedas, o una avalancha que dejara 'impresos' sus dígitos mediante pedruscos que rodaran ladera abajo. Probablemente en algún lugar del universo estén ya grabados los primeros dígitos de Ω. Empero, ningún mortal que descubriera tal tesoro podría verificar su autenticidad, ni tampoco, hacer uso práctico de él".

Cuando esta sección estaba ya en prensa, recibí un telegrama del notorio numerólogo doctor Matrix, donde afirmaba poseer los 31.031.031 primeros dígitos binarios de Ω (que son suficientes, en principio, para contestar un buen montón de preguntas sin interés, además de otras que sí lo tendrían). En la actualidad, el doctor Matrix solicita pujas sobre bits individuales, o bloques de bits consecutivos.

He aquí unos cuantos cabos sueltos relacionados con secciones de meses anteriores.

P. Howard Lyons y otros descubrieron que la mínima rectangulación del rectángulo (mes de agosto) no era única. Los cinco subrectángulos de la segunda solución tienen dimensiones 4 por 8, 1 por 11, 5 por 7, 2 por 6 y 3 por 9.

Abraham Schwartz señaló que el problema "en broma" de dar mate en "una fracción de jugada" (agosto) podría resolverse también alzando, ya la dama, ya el alfil, aproximadamente dos dedos sobre el tablero. Otros lectores me enviaron nuevas interpretaciones de "fracción de jugada": Las blancas coronan un peón en AD8, y piden una dama; las negras han retirado el peón, pero todavía no han colocado la dama pedida en la casilla. Por tanto, sólo hace falta la otra mitad del movimiento para completar el mate. Otro lector hizo notar también que el peón coronado podría haber capturado una pieza negra tanto en AD8 como en CD8.

Mike Jones ha descubierto una disposición de 20 superdamas que no se atacan entre sí (la superdama o amazona reúne los movimientos de dama y caballo) en un tablero de 20 por 20; de esta forma, el mínimo caso no resuelto es ahora el correspondiente a tableros de 21 por 21.

Taller y laboratorio

Llamas en las que el aire se introduce en un gas inflamable en lugar de al contrario

Jearl Walker

Contemplando el familiar espectáculo de la llama producida por un gas al quemarse en el aire, podríamos preguntarnos qué ocurriría si el gas y el aire se intercambiaran, esto es, si se introdujese un chorro de aire en una atmósfera de gas. ¿Podría verse una especie de llama inversa con la misma forma, altura, color y temperatura de la llama normal? Estas cuestiones se le plantearon a Stuart Travis, de Akron, Colorado, quien, con sus investigaciones sobre la llama inversa del aire en una atmósfera de metano, ganó el segundo puesto en la sección de física en la Exposición Internacional de Ciencias e Ingeniería del año pasado. La llama inversa es similar a la normal y sólo muestra algunas diferencias sorprendentes, cuya naturaleza ni Travis ni yo entendemos bien del todo.

Las llamas normal e inversa investigadas por Travis son llamas de difusión. Para que se produzca la combustión debe presentarse una mutua difusión del combustible en el oxidante, y viceversa, en el frente de la llama, zona donde tiene lugar la combustión. En el otro tipo general de llama, llama de gases premezclados, el combustible y el agente oxidante se mezclan antes de que alcancen la ignición, como sucede en los mecheros Bunsen. Las llamas por difusión y las premezcladas pueden ser laminares (esto es, que fluyen suavemente) o turbulentas, en razón de la velocidad de flujo de los gases que intervienen en la combustión. Travis se ha limitado a estudiar las llamas por difusión en régimen laminar. El ejemplo más común de este tipo es la llama de una vela, caso que ya examinamos con detalle en esta sección en el mes de junio de 1978.

Podría pensarse que las llamas laminares por difusión se conocen a fondo, puesto que la llama de las velas ha sido objeto de investigación a lo largo de muchísimo tiempo. No ocurre así. Las llamas son sumamente difíciles de comprender incluso en sus aspectos más generales, no digamos ya en detalle, debido a la amplia gama de fenómenos que en ellas concurren. En su interior tienen lugar una infinidad de reacciones químicas, gran parte de las mismas muy complejas. La luz que emiten se debe tanto a fenómenos químicos como a fenómenos termodinámicos. Los gases se expanden al ser calentados. El calor se transfiere a través de la superficie de la llama de diversas maneras. El flujo tiene tres dimensiones. Además, al investigar sobre la misma llama acabaremos por distorsionarla de forma inexplicable.

Los estudiosos de las llamas se han concentrado sobre algunas de sus características principales: color, forma, distribución de la temperatura y velocidad de la combustión. La velocidad de la combustión es aquella a la cual la llama se propaga en la perpendicular a su frente. La propagación puede ser o bien a través de un gas estacionario o, como sucede en la llama de una estufa, a través de una llama que permanece estacionaria mientras que es el gas el que fluye. Es fácil calcular la velocidad de combustión cuando el frente de la llama es plano; en cambio, cuando el frente es curvo, como ocurre en la mayoría de los mecheros, el tema se complica.

Otra cuestión de interés en el estudio de la llama es la zona de reacción: se trata de la zona donde se mezclan el combustible y el agente oxidante. Parte de la luz que emite una llama se produce por las reacciones químicas que tienen lugar en dicha región. En una llama de difusión normal, la zona de reacción es azul, debido, sobre todo, a las emisiones azules de las partículas excitadas por las reacciones químicas de la región. Si el flujo de combustible se hace suficientemente lento, la llama se torna azul. Ejemplo de este tipo de llama es la de una vela con una mecha muy delgada.

Si aumentamos un tanto el flujo de combustible, comienza a formarse una punta amarilla o blanca en la parte superior de la llama de difusión normal. A medida que crece el nivel de flujo de combustible, aumenta la región blanca o amarillenta, llegando incluso a extenderse por toda la llama relegando a los laterales la zona azul de reacción, de la llama primero, y confinándola finalmente en la región lateral próxima a la base. Las puntas blanco-amarillentas consisten en partículas de carbono que han alcanzado una temperatura tal que les permite emitir radiaciones de forma más o menos continua a lo largo de todo el espectro visible.

La cámara de combustión que Travis empleó en sus investigaciones consistía en un frasco esférico de 500 mililitros colocado boca abajo con la ayuda de un soporte y una abrazadera. El frasco llevaba incorporado a su superficie dos tubos delgados, uno hacia arriba en la parte superior y otro situado lateralmente en el cuello del frasco. El tapón del frasco contenía un tubo que hacía las funciones de mechero. Para conseguir una llama de difusión normal, Travis suministraba metano por el tubo que servía de mechero y aire por el tubo lateral. Para obtener la llama inversa, intercambiaba los suministros. Un pequeño orificio practicado en un lateral del frasco permitía introducir un termopar para medir la temperatura de la llama. A través del tapón, Travis introdujo también dos conductores eléctricos y una resistencia de nicromo que le serviría para iniciar la llama dentro del frasco. Tomaba el aire y el metano de los tanques del laboratorio de su escuela.

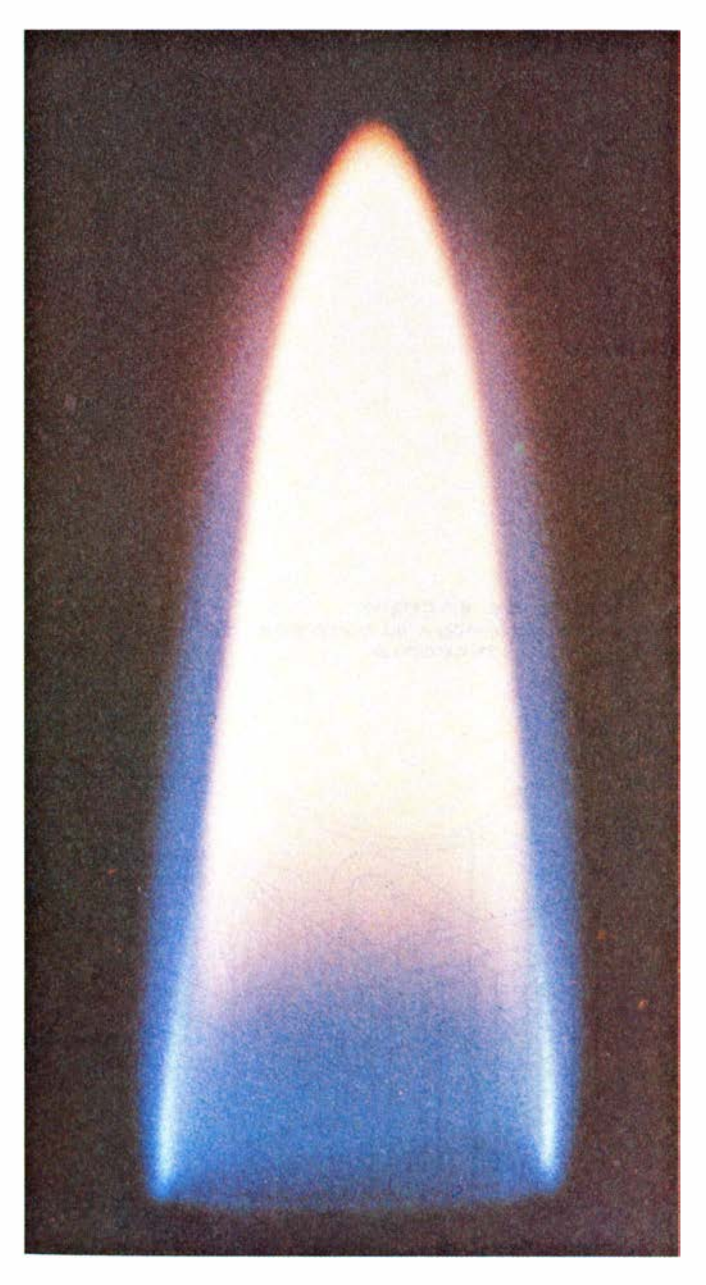
Para encender una llama inversa, Travis colocaba la resistencia de nicromo encima de la boca del tubo que servía de mechero y hacía pasar metano. A los 15 segundos, el frasco estaba lleno de metano y podía encender el gas que salía por el tubo empleado como chimenea de la parte superior del frasco. (Se necesitaban 15 segundos, para evitar que se formara una mezcla con poco de butano y mucho aire que pudiera resultar explosiva). Ajustó el flujo de metano hasta conseguir que la llama, por la chimenea, tuviera unos ocho o diez centímetros de altura. Una vez encendida la

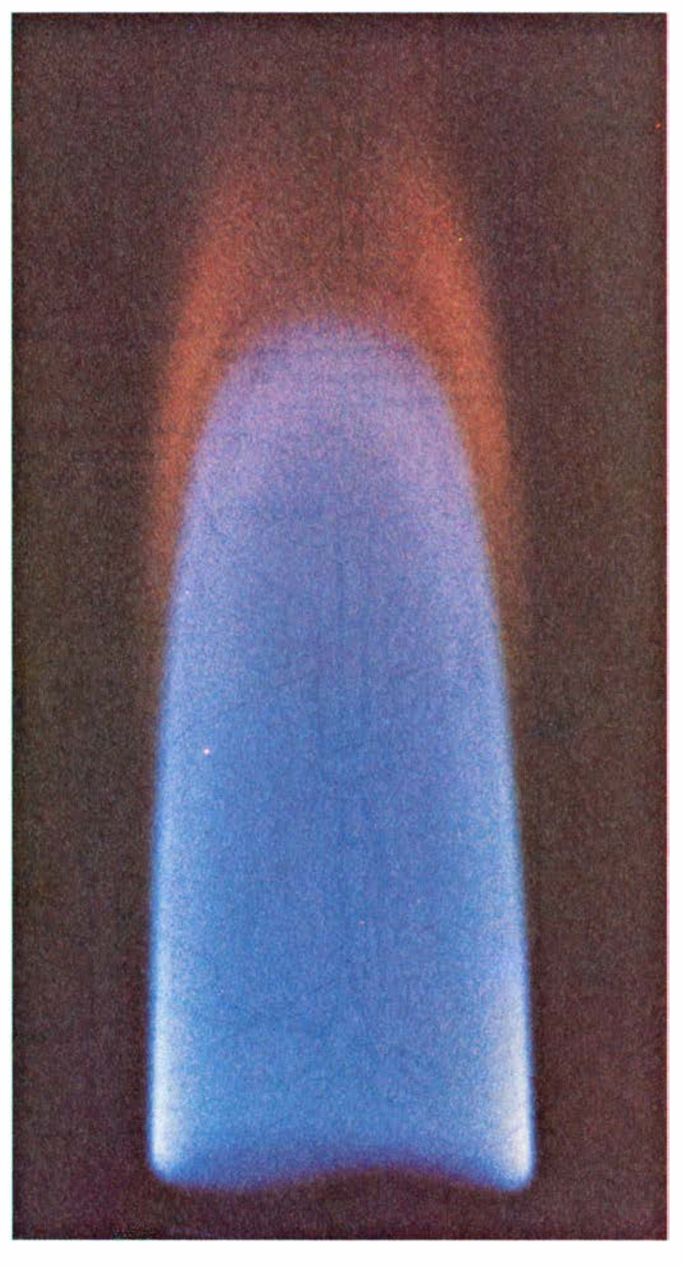
llama inversa dentro de la cámara, la llama de la chimenea quemaba el excedente de metano que salía de la cámara.

El alambre de nicromo estaba conectado, por medio de dos conductores, a una fuente de energía eléctrica que proporcionaba una intensidad suficiente para ponerlo al rojo. A continuación hacía pasar un pequeño flujo de aire (aproximadamente seis litros por minuto). Al principio, la cantidad de aire tenía que ser baja, ya que una gran cantidad de oxígeno podía haber provocado la explosión del frasco. El alambre al rojo prendía la llama en la boca del mechero, con lo que se obtenía la llama inversa. A partir de ese momento, se desconectaban los hilos de la corriente y, moviéndolos desde fuera, los hacía caer de su posición inicial en la boca del mechero. La llama se distinguía mejor en la oscuridad; por eso, la mayoría de estas operaciones las hacía en una situación de penumbra. Para apagar la llama, Travis cerraba primero la entrada de aire con el fin de evitar cualquier peligro de explosión.

Nuestro físico encendía la llama normal con un procedimiento parecido. De nuevo, ponía mucho cuidado en evitar encender la llama en la parte superior del frasco en tanto hubiera en el interior del mismo mucho aire y poco metano. Llenaba, primero, el frasco del metano; a continuación, abría la tubería del aire. Antes de apagar la llama, cerraba el paso del aire. Cuando le pregunté sobre el peligro de explosión, me contestó que no había tenido problemas con la llama inversa pero que, con la llama normal, muy de vez en cuando sufría una explosión que le hacía saltar el tapón del frasco. Para evitar un accidente más importante, quien se apreste a, o quiera realizar, un experimento de este tipo debería emplear una careta de protección. Además no conviene usar como cámara de combustión un frasco mayor de 500 mililitros, pues entonces el riesgo de explosión se elevaría.

La distribución de la temperatura en la llama se medía con una sonda termopar, introducida en la cámara de combustión a través de un pequeño agujero. Lijó la punta del alambre del termopar, lo soldó y lo volvió a lijar para que fuese lo más pequeño posible; aún así, la





Llama normal **(izquierda)** *y llama inversa* **(derecha)**

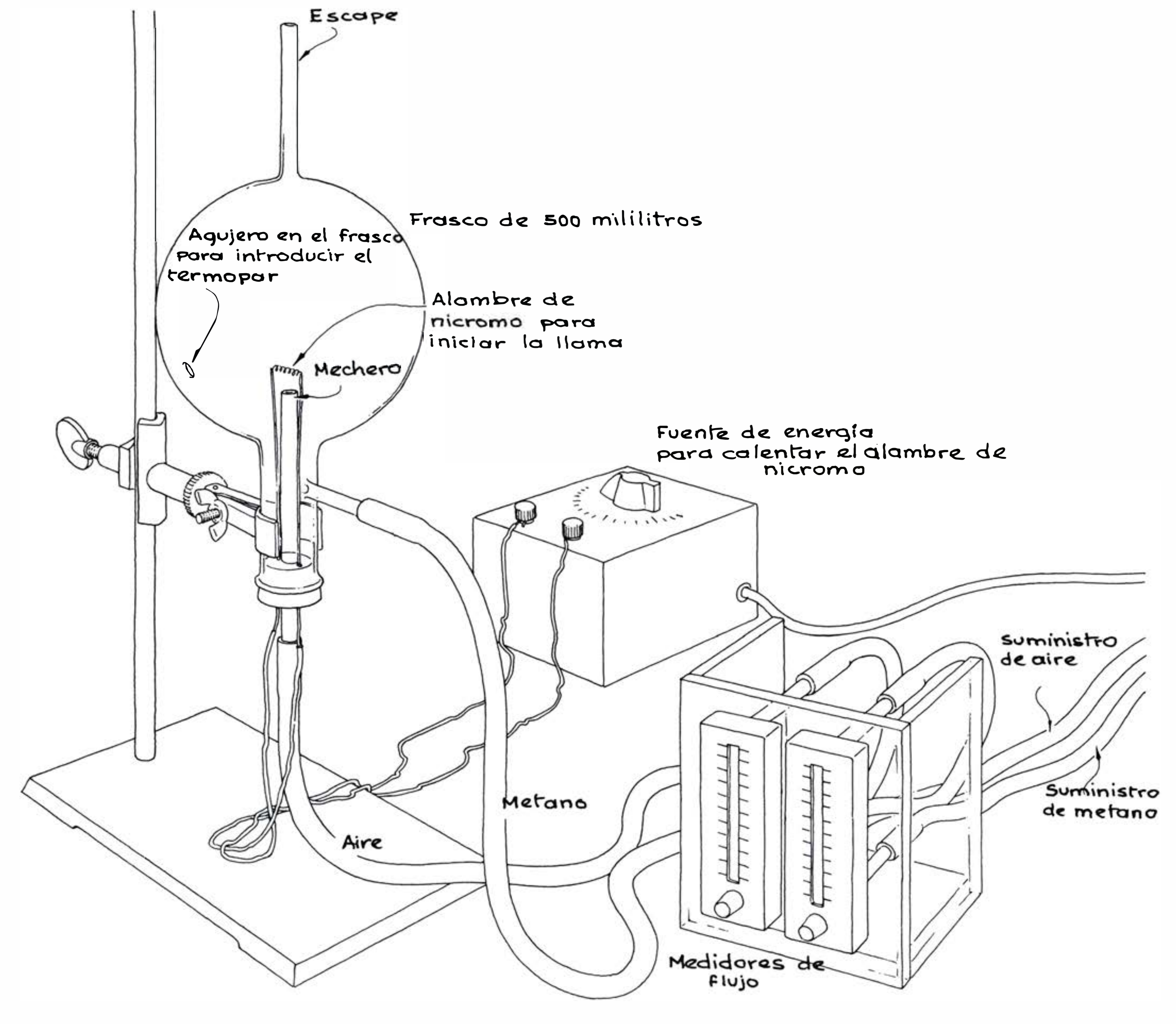
sonda era relativamente grande comparada con la llama y por eso sólo podía medir temperaturas medias. También se debe tener presente dos efectos incuantificables de la sonda: la distorsión de la llama y el enfriamiento que produce y, por tanto, la distorsión que provoca en la distribución de las temperaturas.

Para conocer la posición de la sonda en la llama, Travis montó una regla por fuera de la cámara de combustión de forma que la distancia desde el extremo del termopar al agujero por dentro de la cámara fuera igual a la distancia entre el agujero y la regla; y así, cuando el extremo se desplazaba cierta distancia dentro de la llama, el alambre rígido del termopar se movía un recorrido igual por la regla, con lo cual le era posible medir el desplazamiento en la llama.

Travis calculaba el flujo de gas con dos medidores de flujo relativamente baratos, que trabajaban en el rango de unos pocos litros por minuto. Si los flujos eran muy pequeños, medía de forma aproximada el caudal haciendo burbujear el gas en un recipiente con agua. Contando el número de burbujas por minuto que pasaban a través del agua, y estimando el tamaño medio de las mismas, podía con un simple cálculo conocer el volumen del flujo.

La configuración esférica de la cámara de combustión distorsionaba las fotografías de la llama. Por ello, Travis sustituyó la cámara esférica por otra rectangular hecha con placas de cristal común. Cuando encendía la llama, debía trabajar rápidamente puesto que el calor enseguida rompía el cristal. Evidentemente, podía haber utilizado un cristal más caro, capaz de soportar las tensiones térmicas y entonces se hubiera ahorrado el problema. Para sacar las fotos, Travis empleaba una cámara de 35 milímetros con una lente de 55 milímetros y un anillo de acercamiento. Hizo las fotografías en blanco y negro con una película Kodak Tri-X, una apertura de diafragma de 2,8 y una velocidad de exposición de 1/30 de segundo.

Para medir la velocidad de combustión, el joven físico empleó la técnica conocida como método de Gouy, basado en el área total del frente de la llama. La velocidad de combustión se supone constante en todo el frente de la llama a pesar de la curvatura del mismo. El caudal de gas del mechero es igual al área de la boca del mechero multiplicada por la velocidad del gas en dicha boca (este resultado ha de coincidir con el valor leído en el medidor de flujo). Por ser estacionaria la llama, el caudal de gas debe equivaler también a la velocidad de combustión multiplicada por la superfi-



Cámara de combustión ideada por Stuart Travis para la llama inversa

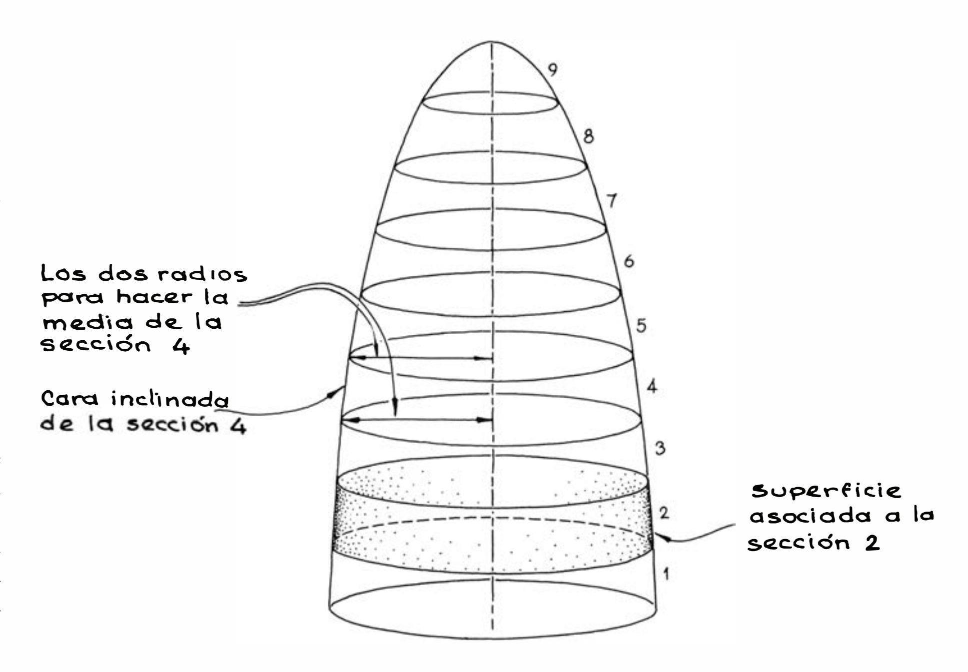
cie total del frente de la llama. Si la superficie se puede calcular, sólo nos queda como incógnita la velocidad de combustión.

Para determinar la superficie total del frente de la llama, hay que fotografiar la llama y, en la foto, dividir la llama en dos partes iguales por medio de un eje central vertical. Una de las mitades se divide en varias secciones como se indica en la ilustración de la derecha. Dos de los lados de cada una de estas secciones corresponden a los radios desde el eje central hasta el frente de la llama y otra cara es la zona inclinada del frente de llama. La superficie a lo largo de la circunferencia de la llama correspondiente a una de estas secciones se calcula multiplicando el número pi por la media de los dos radios de la sección y por la longitud tomada a lo largo de la superficie inclinada de la llama. El área total obtenida viene a ser el área total del frente de llama.

¿Cuál es la exactitud de este cálculo? El cálculo puede resultar falseado por causas varias. Una es que determinar la posición del frente de la llama puede ser difícil si la zona de reacción es muy ancha, como es normal que ocurra en llamas con baja velocidad de quemado. Otra fuente de error proviene de la dificultad de calcular la superficie en la parte superior de la llama, donde el frente de la misma puede estar sustancialmente curvado. Nos puede inducir a cometer un error de bulto el que la base de la llama esté definida pobremente en la fotografía. Y, finalmente, la velocidad de combustión de la llama puede venir distorsionada por la expansión térmica de los gases cuando entran en la zona de reacción y se calientan.

Predecir la forma o altura de una llama de difusión normal tampoco es grano de anís. Ambas cuestiones dependen de una compleja interacción de la transferencia de calor, de la difusión y la velocidad de las reacciones químicas. En un análisis muy simple podría decirse que la altura de una llama de difusión sería directamente proporcional al caudal de gases e inversamente proporcional a la velocidad media de difusión de los mismos.

Las llamas inversas hechas por Travis en su aparato son tan difíciles de analizar como las llamas normales, sino más. Ambos tipos de llamas difieren notablemente en tamaño, y color. Una llama inversa suele tener la parte superior redondeada, no cónica, que es lo que ocurre en una llama normal. Una llama normal con un determinado caudal de metano es mucho más alta que otra inversa con el mismo caudal de aire. La llama normal suele tener una base poco definida, mientras que la llama inversa ostenta una base precisa y definida; además, dicha base está separada de la boca del mechero por un espacio relativamente ancho y muerto.



Cómo seccionar la llama para calcular su superficie

La razón de diferencia más destacada entre ambas llamas es su color. La normal, producida con un caudal de combustible pequeño, es totalmente azul. A medida que aumentamos el caudal de metano aparece en la parte superior una punta blanca o amarilla y las partes azules se reducen y desplazan hacia la base. Una llama inversa es totalmente azul hasta que se introduce en ella una gran cantidad de aire; a partir de entonces, la punta se torna naranja.

La distribución de temperaturas difiere ligeramente en una y otra clase de llama. La llama normal tiende a mostrar su parte más caliente a los lados del cono interior obscuro; la inversa tiene algunas zonas más calientes por encima del cono interior. Hay que advertir que estas medidas están afectadas por la introducción de la sonda del termopar en la llama y porque el termopar promedia las temperaturas de una región, aunque dicha región sea pequeña.

La llama inversa tiene también una velocidad de combustión más alta. Cuando Trevis ajustó una llama normal y una inversa de forma que alcanzaran la misma altura, la llama inversa conseguía alrededor del doble de velocidad de combustión que la normal. Cuando ajustó las llamas a iguales caudales (metano para la normal y aire para la inversa) la relación se acercó a 10. En ambos tipos de llama, la velocidad de combustión descendía (desde unos pocos centímetros por segundo) al aumentar el caudal, ya que el área de la parte anterior de la llama no aumentaba tanto como el caudal.

¿Por qué la llama inversa es diferente de la llama normal? No estoy seguro del todo, pues ni siquiera acaba de conocerse bien el mecanismo de la llama normal de difusión. La llama inversa y la normal se parecen cuando los caudales son bajos. En esa condición, una y otra tienen forma redondeada en la parte superior y color azul en la zona de reacción. Al aumentar el caudal de metano en la llama normal, aparece rápidamente la punta amarilla o blanca. La correspondiente punta naranja de la llama inversa se desarrolla menos deprisa ya que el caudal de aire debe aumentarse considerablemente para provocar su aparición.

No deja de desconcertarnos que el color de la punta de ambas llamas no sea el mismo. Debe ser (por lo que yo sé) a causa de que en una y otra el color de la punta obedece a las emisiones visibles de partículas de carbono incandescentes. Puesto que la alimentación de aire en la llama inversa ha de ser relativamente alta para provocar la punta naranja, la formación de partículas de carbono incandescente podría requerir también una alimentación de aire relativamente alta.

El hecho de que el metano se difunda desde la parte exterior de la llama inversa, en lugar de hacerlo desde el interior, como sucede en la llama normal, podría constituir la clave de las diferencias entre los dos tipos. El metano nece-

sita para su descomposición una temperatura apreciablemente alta. Si este gas proviene de la parte interior de la llama, penetra en las partes más calientes de la misma muy rápidamente. No se conocen los mecanismos por los cuales se terminan formando partículas sólidas de carbono, aunque se sospecha que el proceso no queda completado hasta que los gases alcanzan la parte superior de la llama. Cuando el metano se difunde en una llama inversa, entra por la parte más fría y se calienta gradualmente; ello justifica que las reacciones que forman las partículas puedan ser diferentes. Sin embargo, mis especulaciones en estas cuestiones son puras suposiciones en espera de que se hagan más trabajos sobre las llamas inversas.

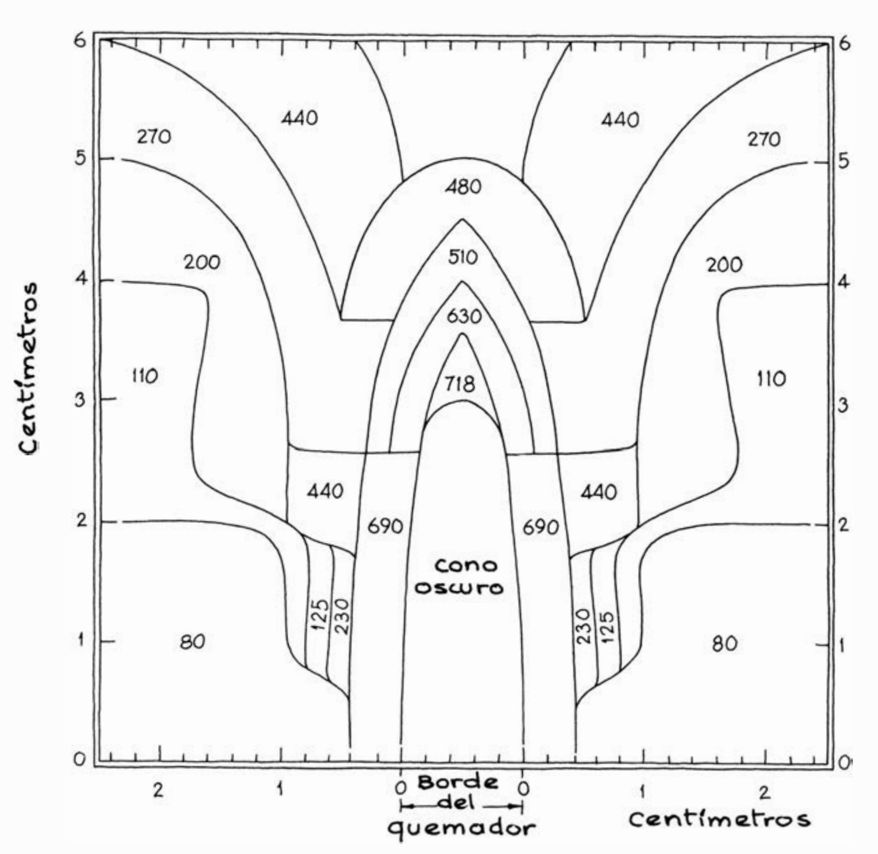
Temperaturas (en grados Celsius) en torno a una llama normal

Una llama normal de difusión que sea estable se asentará ligeramente encima de la abertura de un quemador a una altura en que la velocidad de combustión iguala la velocidad del gas. Si una de las velocidades cambia de una forma significativa, la llama se moverá, ya sea alejándose de la boca o introduciéndose en ella. La velocidad del gas disminuye algo, a medida que la corriente del mismo se difunde desde la boca del mechero, de suerte que el frente de la llama se asiente en aquel punto por encima de la boca en que ambas velocidades se igualan. Algunas llamas (como las de chorros de etileno) pueden elevarse bastante por encima del quemador y estabilizarse por algún tiempo. No es fácil obtener este tipo de llamas elevadas, pues requieren un equilibrio perfecto entre la velocidad del gas y la velocidad de combustión, para el experimentador.

Travis no logró que su llama normal se levantara de esta forma, pero lo consiguió fácilmente con la llama inversa. Cerraba la alimentación de metano y rápidamente la volvía a abrir. La llama se elevaba varios centímetros por encima de la boca del quemador, permaneciendo así hasta cuatro minutos. Aunque no estoy seguro de dónde estriba la razón de ese comportamiento de la llama inversa, sospecho que quizá la relativamente alta velocidad de combustión pudiera ser la clave.

Cuando en una llama normal se abre mucho la alimentación de metano, ésta se vuelve turbulenta. La llama inversa no manifiesta tal turbulencia. Al abrir mucho su alimentación, la llama se aleja de la boca del mechero y se apaga.

No hallé mención alguna de llamas inversas en la bibliografía pertinente. Quizá sea Travis la primera persona que las ha investigado. Si se desea proseguir su estudio, habrá que insistir en la experimentación. Intentar, por ejemplo, con otros hidrocarburos gaseosos distintos del metano (se debe tener precaución si son tóxicos o altamente explosivos). Travis ensayó también con propano. Se puede empezar por ahí. En mi anterior artículo sobre la llama de una vela, descubrí de qué manera un simple espectroscopio puede servir para observar las inversiones de los radicales químicos CH y C_2 de las zonas de reacción. Se podrían hacer observaciones similares en las zonas de reacción de una llama inversa. Si no se observan emisiones, los procesos químicos de las zonas de reac-



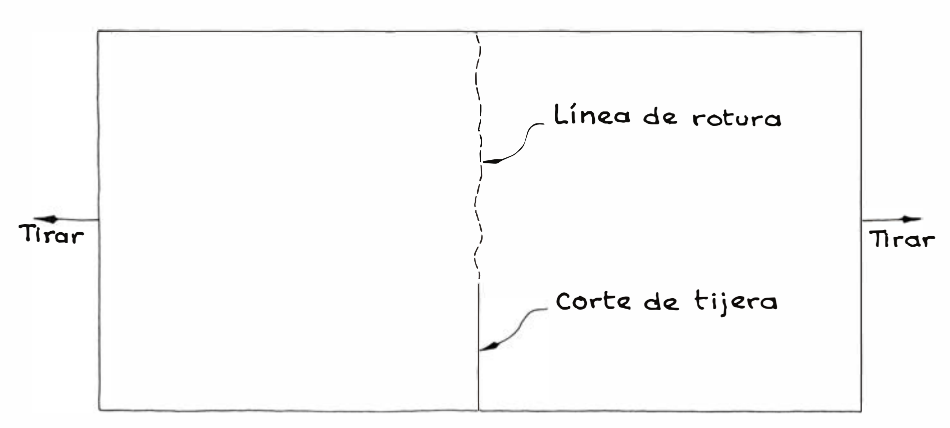
Temperaturas en torno a una llama inversa

ción en ambos tipos de llama serían muy diferentes.

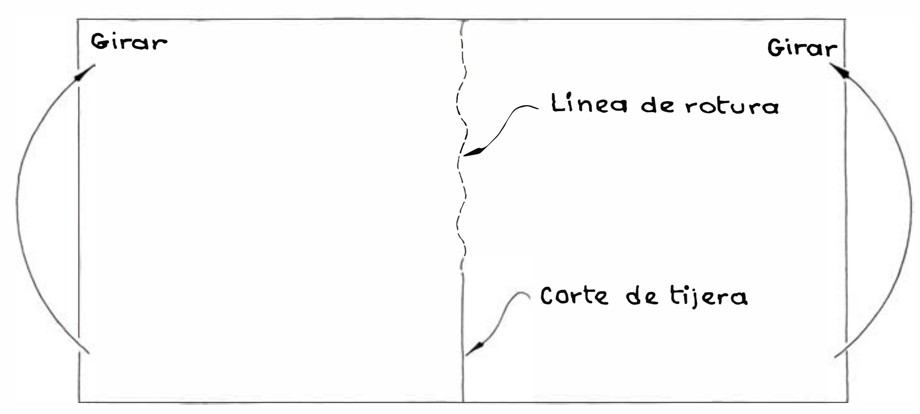
He recibido copiosa correspondencia a propósito de la similitud entre la caída de árboles y chimeneas, que se discutió en esta sección en abril. James D. Plimpton, de Albuquerque, señalaba que un árbol con muchas hojas, en su caída, podía "flotar" brevemente un instante antes de tocar el suelo, debido a la resistencia del aire contra sus ramas. Un aspecto más peligroso de la caída de árboles lo describe Paul R. Burnett, de Temple Hills. A un árbol que va a ser derribado, se le hace una gran muesca en un lado y después un simple corte horizontal en el lado opuesto. Cuando empieza a caer, bascula sobre los dos dedos de madera que quedan entre la muesca y el corte opuesto. Si el árbol es joven, puede combarse apreciablemente por la resistencia del aire según cae; un árbol vivo nunca se quebrará en razón de la gran resistencia a la tensión de la madera verde. Sin embargo, la parte de madera sobre la que bascula, se rompe, lanzando el extremo del árbol hacia arriba con una fuerza considerable. Ese extremo puede moverse entonces en la dirección de caída del árbol, como la base de una chimenea que se derrumba. Si el árbol tiene muchas ramas gruesas, puede almacenarse en ellas mucha energía cuando se arquean al golpear el suelo. Las ramas pueden impulsar bruscamente el tronco hacia el tocón con gran fuerza. Burnett ha visto tocones casi arrancados del suelo por el impacto. Este repentino y violento retroceso es particularmente peligroso para quien está junto al tocón o incluso cerca de él, después de la caída del árbol.

El comportamiento en la caída de un árbol muerto y parcialmente podrido es aún más curioso. John D. Engels, de North Bend, ha observado que la rotura de estos árboles al caer es muy similar a la de las chimeneas. Los árboles muertos y podridos se rompen a veces por la mitad; los dos trozos caen a ambos lados del tocón, siendo un grave dilema para el leñador decidir hacia qué lado correr. Engels ha visto también arquearse al caer árboles con muchas hojas verdes. Algunas veces el árbol se endereza bruscamente al acercarse al suelo, de manera que la mitad superior alcanza el suelo antes que la inferior.

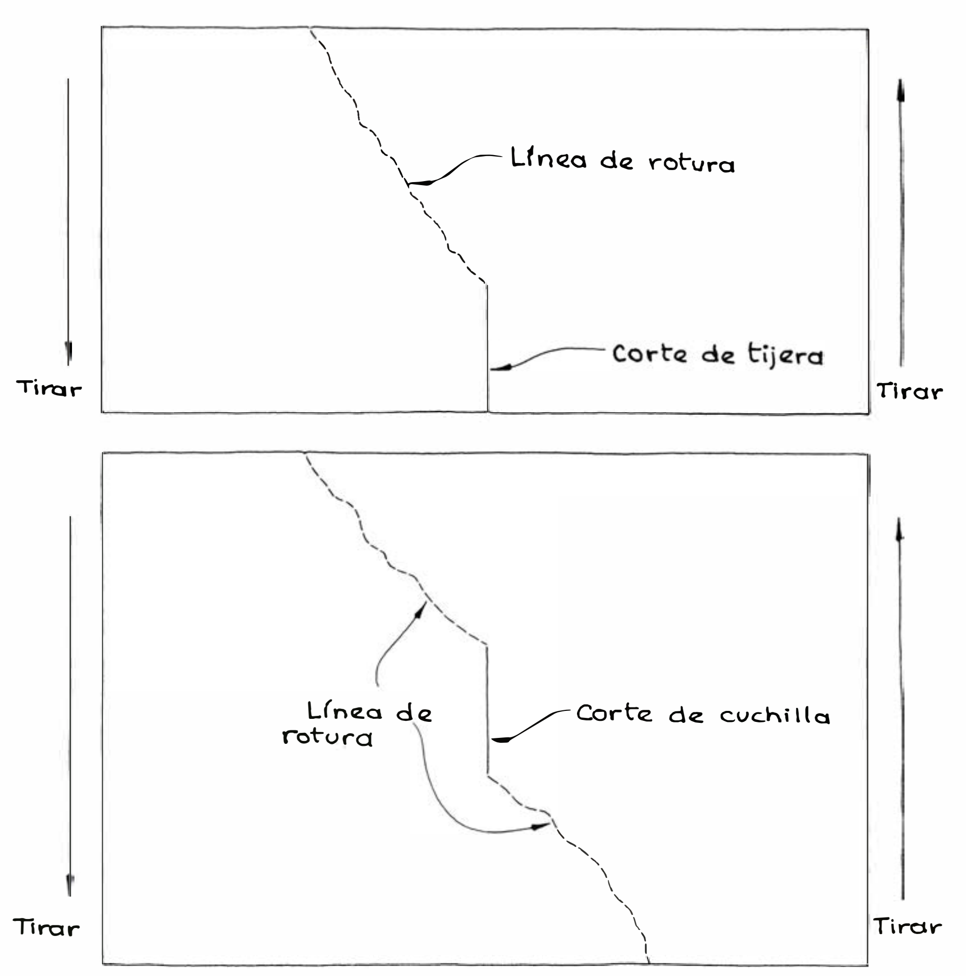
Después de leer mi descripción (en el mismo artículo) de cómo se rompen las puntas de los lápices, Gerard R. Martin, de Napa, me remitió la descripción de un extraño efecto que habían observado unos alumnos de química con agitadores de cristal. Si una de estas varillas



Modo de fractura mixto. Modo I: Esfuerzo puro de tensión



Modo I: Esfuerzo puro de pandeo



Modo II: Propagación de una rotura por cizalla

caía, se rompía frecuentemente en tres trozos de longitudes iguales y unas pocas astillas. Este modelo de rotura fue apellidado "regla de pi", pues las varillas se rompían aproximadamente en 3,1 partes. Martin y otros matemáticos generalizaron el resultado (como suelen hacer los matemáticos) incluyendo barras de tiza. ¿Existían otros ejemplos de la regla de pi?

Comprobé esta regla dejando caer trozos de tiza en el suelo, después de marcarlos de manera que pudiera reconstruirlos y situar el punto de impacto. Aunque la regla no se observaba con exactitud milimétrica, la tiza solía romperse en tres partes (rara vez de la misma longitud) con fracturas similares a las de las minas de lápiz. Mi trabajo obtuvo distintos comentarios de mis colegas de la Universidad del estado de Cleveland. Karl Casper descubrió que si se reduce la longitud de la tiza en un tercio, se romperá por lo general en dos partes, y no en tres. Bien podría suceder que una pieza de tiza se rompiera primero a un tercio aproximadamente de su longitud total y los restantes dos tercios lo hicieran aproximadamente por la mitad, cuando la parte superior golpeara el suelo un instante después. Se puede intentar hacer fotografías de alta velocidad para averiguar si mi elucubración está fundada o si la pieza entera de tiza vibra y se rompe por dos sitios a la vez.

Otro ejemplo de fracturas, que en principio parece no tener relación con los anteriores, me lo mandó J.G. Nandris, de la Universidad de Londres. Alrededor del año 2000 a.C. se erigió un menhir (o monolito vertical) enorme, Le Grand Menhir Brisé, en Carnar. La piedra medía 21 metros de altura y pesaba más de 300 toneladas. No se comprende cómo hombres con herramientas primitivas pudieron construirlo y levantarlo. La piedra permaneció de pie hasta hace cerca de 700 años, en que se vino abajo y se partió en cuatro trozos. La mitad superior descansa actualmente en tres partes alineadas hacia el Noroeste, después de haber caído, al parecer, en una dirección casi opuesta a la tomada por la mitad superior.

¿Por qué cayó la piedra? ¿La partió un rayo o fue derribada por los indígenas? ¿Por qué cayeron las dos mitades en direcciones opuestas? Lo más probable es que la derribara un terremoto que la cortó por la mitad. Cuando la parte superior cayó al suelo se rompió en tres trozos, de forma semejante a las barras de tiza que yo tiraba. Las superficies de las roturas de las cuatro piezas recuerdan las porciones de las tizas. La mitad inferior de la piedra continuó sobre el trémulo suelo (si realmente fue derribada por un terremoto) hasta que también cayó, pero lo hizo en una dirección diferente de la otra mitad. Tanto el menhir como el terremoto que lo destrozó debieron ser pavorosos.

El tipo de fractura común en las minas de lápiz, chimeneas, tizas e incluso el del menhir, se denomina fractura de tipo mixto, pues incluye dos modos generales de fractura, según me explicó recientemente Anthony R. Ingraffea, de la Universidad de Cornell. Las ilustraciones de la página anterior enseñan cómo cortar y rasgar hojas de papel para demostrar esos modos. Se origina una fractura del tipo llamado modo I, cuando el material se somete a esfuerzos puros de pandeo o de tensión, ninguno de los cuales puede ser causa de que los dos trozos resultantes de la fractura deslicen uno sobre el otro. Para demostrar este fenómeno no hay más que hacer un corte en una hoja de papel y tirar o hacer girar los extremos según se indica en las dos ilustraciones superiores de esa página. La rotura se propaga en una línea recta, paralela al corte realizado.

Si deslizamos, una sobre otra, ambas fracciones del corte, sin que lleguen a separarse, la rotura pertenece al modo II. Córtese las hojas de papel según se indica en la ilustración inferior de la página y tírese en sentidos contrarios, según convenga. La rotura esta vez se propaga de una forma curvilínea.

Las roturas que se propagan a lo ancho de una chimenea que cae o de una mina de lápiz bajo esfuerzo son del tipo mixto: pandeo (modo I) y cizalla (modo II). Por tanto, la rotura muestra algo de ambos modos de fractura. La dirección en la cual la rotura se curva durante la última parte de la propagación es diferente en las chimeneas y en las minas de lápiz, debido a la dirección de la cizalladura. Por ejemplo, si la chimenea cae hacia la izquierda, la cizalladura en sentido de las agujas del reloj hace que la rotura se curve hacia abajo, según pasa a lo ancho de la chimenea. En una mina de lápiz la dirección de cizalladura desvía la curva de la rotura hacia arriba, lejos del extremo.

Había un error en la ilustración del circuito de amplificación del sismógrafo descrito en esta sección en septiembre. La línea del condensador de 0,4 microfarad al terminal numerado "2" en la segunda puerta debe conectarse a la línea entre la resistencia de un kiloohm y las resistencias de 100 kiloohm.

Libros

Microeléctronica, la cabaña montañesa y una interesante histotecnología ilustrada

Antonio Alabau, Clemente Sánchez-Garnica y Mercedes Durfort

MICROPROCESADORES Y MICROORDENADORES, por R. Lyon-Caen y J. M. Crozet, Editorial Toray-Masson, S.A., 1979, 171 páginas. Se trata de la traducción al castellano del libro "Microprocesseurs et Microordinateurs", publicado por la editorial Masson en 1976. En este primer párrafo hemos querido resumir los tres aspectos de esta obra que nos interesan destacar, y que especifico: se trata de un libro de microordenadores, publicado en 1976 y traducido al castellano.

La actualidad del tema "microcomputador" es innegable. (Recuerde el lector el número monográfico que *Investigación y Ciencia,* dedicó, en noviembre de 1977, al tema.) Este tipo de elementos, aparecido tímidamente en 1971, se ha situado entre los componentes electrónicos con mayores posibilidades, en cuyo entorno se manejan cifras de negocios espectaculares.

A raíz de los avances en los procesos de fabricación de circuitos integrados y debido sobre todo a la reducción de sus costos por su producción en cantidades elevadas, la utilización de los microcomputadores, bien como sistemas autosuficientes dedicados a funciones de computación, bien como elementos de equipos electrónicos más sofisticados, aumenta de día en día, incluso en países como el nuestro. El número de personas implicadas en dicho fenómeno se incrementa y en consecuencia las necesidades de formación de personal técnico especializado en su utilización crece, y con ella, la de textos que faciliten dicho aprendizaje. Ello, lógicamente, lleva a que las editoriales incluyan en sus catálogos libros sobre estos temas, cuya venta está casi asegurada. La calidad, orientación y contenido de dichos textos son ya cuestiones diferentes.

Los principios que inspiran la arquitectura de un microcomputador son, fundamentalmente, los mismos en los que se basa la de cualquier computador, por tratarse en todos los casos de máquinas universales de programa almacenado. Las características particulares de los microcomputadores, derivadas del hecho de haber sido fabricados en unos pocos módulos de circuito integrado (a veces en uno único), sin ser menospreciables, están desde luego supeditadas a los anteriores, y en cualquier caso, si jerarquizamos el estudio de los microcomputadores, aparecen en los niveles inferiores.

Con esto queremos decir que, al estudiar los microcomputadores, no es aconsejable mezclar conceptos situados en niveles distintos; ni mucho menos hay que olvidar el enorme bagaje de conocimientos que, bajo la denominación de ciencias de los computadores, ha ido acumulándose desde finales de la década de los 40.

Con los microcomputadores, no se inventa apenas nada en el campo de la arquitectura de computadores. Tan sólo se pone al alcance de muchos lo que antaño estaba reservado a pocos. En nuestra opinión, la formación de un especialista en microcomputadores (que habrá de ser un especialista en sistemas microinformáticos), para que resulte estable, debe descansar sobre tres puntos fundamentales: conocimientos de arquitectura de computadores, conocimientos de programación (y de buena programación) y conocimientos de electrónica digital; éstos, en último lugar y con mucha menos importancia que los anteriores.

Los autores tratan sumariamente todos los puntos precedentes, pero sin centrarse en ninguno de ellos. Un rápido repaso al índice de materias nos dará más información. El capítulo primero contiene una introducción a los microprocesadores y a los microordenadores, sorprendentemente corta para nuestro gusto.

El segundo capítulo quiere ser una descripción de los principios básicos de funcionamiento de un microprocesador. De una forma bastante desafortunada, se presenta el microprocesador como una generalización de los sistemas lógicos especializados; sigue una explicación de la arquitectura de un microprocesador a nivel de transferencia de información entre registros, para finalizar con un intento de descripción del funcionamiento y aplicación de algunos elementos, indebidamente llamados complementarios. Este capítulo es muy flojo en lo concerniente al tratamiento de la temas básicos de arquitectura de los computadores.

El capítulo tercero estudia algunos detalles sobre la lógica de control de un microprocesador (o como dicen los autores: "... demostrar cómo la unidad de control genera las salidas en función de las diferentes entradas"). Se pretende ilustrar detalles sobre la ejecución de las instrucciones e introducir conceptos sobre arquitectura de sistemas microprogramables, pero la exposición es bastante confusa. En particular, en un momento en que el lector no tiene todavía demasiado claros ni la arquitectura ni el funcionamiento de un computador a nivel de transferencia de información entre registros, se le obliga a descender a la arquitectura y funcionamiento de la máquina a nivel de microprogramas, informaciones, en muchos casos, innecesarias para la utilización normal de un microcomputador. Otra fuente de confusión es la inclusión allí del estudio de los sistemas a "rebanadas" (bit-slice) que, por su naturaleza, constituyen otra familia de sistemas distinta de los computadores.

En el capítulo siguiente se aborda el problema de realización de un microordenador basado en un microprocesador. Aunque intenta cubrir el conjunto de elementos a tener en cuenta en una tal realización, nos parece que el enfoque es bastante desafortunado; en particular, en el tratamiento del tema de la comunicación con el exterior y, muy especialmente, en la elección del ejemplo para ilustrar tales conceptos. La utilización de los conocimientos anteriormente ex-

puestos, aplicados a un ejemplo, constituye el sujeto de estudio del capítulo siguiente. Esquema, a nuestro aviso, parcial. Tampoco estamos de acuerdo con la elección del microprocesador utilizado, que es simple como apuntan los autores, pero obsoleto ya en el momento en que redactaron la obra. El capítulo sexto, que se adivina, va dedicado al estudio de la programación. El propio título nos sume ya en la perplejidad: "Programación sobre microordenador lógico de ayuda a la puesta a punto"; quizás el lector descifre el significado. La exposición se reduce a un conjunto de explicaciones de conceptos, que no definiciones, y poco más. Por lo que se refiere al contenido, pues, el libro resulta bastante deficiente y muy poco estructurado.

El hecho de que fuera escrito en 1976 queda reflejado en el contenido del séptimo capítulo: "Ensayo de clasificación de los principales microprocesadores actuales", que de vigentes no les queda nada. El contenido está sencillamente preterido y ampliamente superado por los sistemas *actuales*. Finalmente, por lo que se refiere al estilo de origen se adivina que fue muy poco riguroso. El francés coloquial utilizado, traducido casi literalmente, produce resultados que sorprende verlos impresos; frases como "*difícil de poner en obra*", "*datos intermediarios*" o incluso "*microprocesadores más íntegros*", salpican el texto.

Hay, no obstante, un concepto que puede ser fuente de grave confusión para el lector. Se trata de la traducción de "software" (en inglés) o "logiciel" (en francés), traducido a la ligera por *lógica*, término que encierra un claro significado en este campo. Así, por ejemplo, se habla de *lógica de control*, en el que su significado es el habitual y equivalente a circuitos lógicos de control y *lógica de ayuda a la puesta a punto* refiriéndose al "software" (soporte lógico, hubiera podido ser una traducción aceptable, que, además, es la que goza de mayor uso como saben los lectores de la revista). En resumen, se trata a nuestro entender de un libro flojo en su contenido, extemporáneo en su publicación y poco cuidado en su traducción, que se suma a los escasos textos que sobre el tema han sido publicados en castellano. (A. A.)

PASADO Y PRESENTE DE LAS RAZAS VACUNAS SANTANDERINAS DE MONTAÑA, por Angel de Miguel Palomino. Instituto de Estudios Agropecuarios, Diputación Provincial de Santander. El trabajo es simpático, porque en él se realiza una selección de textos y comentarios en relación con las razas vacunas santanderinas, particularmente en lo que se refiere a la tudanca, citando también datos de otras vacas tan interesantes como la pasiega, lebaniega y campurriana, desaparecidas ya para siempre, hecho que debiera producir sonrojo a todos los que, por acción u omisión, fueron responsables de tamaño desaguisado. Los españoles tenemos una peculiar tendencia a destruir nuestros propios bienes, ya sean paisajes, monumentos, animales o ambientes. Cuando hemos logrado nuestro propósito, nos cubrimos de luto y pasamos la vida arrastrando la pena por aquello que perdimos. Cuando yo llegué a Zaragoza, observé que estas buenas gentes guardan un recuerdo grabado en sus genes. Se trata de la que se llamó "Torre Nueva", un maravilloso ejemplar de estilo mudéjar, que fue derribado por la estupidez colectiva. Decían que se podía caer porque estaba un poco inclinada. Según parece, se comprobó luego que se había construido así. Pero el mal ya estaba hecho, y acciones similares se siguen produciendo.

He hecho este inciso porque la actividad contra nuestras razas animales autóctonas nos lleva a situaciones muy parecidas. En muchas regiones españolas, aunque yo diría que en todas, las peculiaridades del suelo, latitud, climatología, etcétera, motivaron la aparición de ejemplares animales únicos, con unas características comunes: su belleza, rusticidad y resistencia a las enfermedades. Baste recordar la oveja merina, la gallina castellana negra, el caballo español o el toro de lidia.

Curiosamente, por esos caprichos de la naturaleza, esta piel de toro que nos sirve de suelo patrio, ha resultado pródiga en estirpes bovinas de calidad indiscutible. En una rápida panorámica podríamos echar una mirada a la retinta andaluza, blanca cacereña, negra avileña, morucha salmantina, rubia gallega, asturiana, pirenaica, murciana, etcétera. Por lo que se refiere a Santander, destacan la levaniega, la pasiega, la campurriana y la tudanca. Pues bien, el papanatismo nacional y la incompetencia radical de quienes fueron responsables de conservar la cabaña ganadera han destruido la mayor parte de los maravillosos ejemplares que hemos relacionado más arriba. Y todo ello sin intentar siquiera un experimento de selección, para lograr, con nuestras propias materias primas, resultados que, sin ninguna duda, hubieran superado a los mediocres que nos van proporcionando las razas importadas. Se trajeron gallinas de América para sustituir a las nuestras. Se importaron razas porcinas, ovinas y hasta vacunas, que no se adaptan a las peculiaridades de nuestro clima. Se pretende con ello mejorar lo que teníamos, pero sólo se logra, aun cuando sea utilizando el "concepto de heterosis o vigor del híbrido", disminuir la capacidad de resistencia de los animales, la difusión de enfermedades exóticas y el incremento de producción a costa de la calidad y de la "naturalidad". Vacas que difícilmente alcanzan los 10 años de vida. Aves que padecen todas las virosis que se han inventado. Carnes insípidas. Leches sin grasa, brucelosis, tuberculosis, enfermedad de las mucosas, rinotraqueítis, etcétera. Es lo que llamamos hoy patología de las colectividades ganaderas. Pero no hay que echar la culpa al sistema de explotación. El que enferma es el animal vivo. Sólo padece una enfermedad quien está predispuesto para ella. Nuestras razas autóctonas, lo mismo que los hombres de nuestras tierras, sabían vivir en España.

Los mozos de hace 30 años daban un porcentaje muy grande de "cortos de talla". Cuando han comido adecuadamente, cuando sus condiciones higiénicas han mejorado, toda la potencia de nuestras razas se ha actualizado. Ahí están nuestros mozos de hoy: altos, esbeltos y hermosos, sin tener que envidiar para nada a las tradicionales razas nórdicas. ¿Qué hubiera pasado si se hubiera actuado de modo similar con nuestros animales?

No he tenido el placer de conocer a las vacas levaniega, pasiega o campurriana. Los viejos del lugar, mis vaqueros de Madrid, me contaban que eran unos animales de capas llamativas, cornamenta singular, resistencia incomparable y rusticidad sólo pareja a la de los hombres que las manejaban. Sin probar un gramo de piensos compuestos, sólo con la hierba de los prados montañeses, aun cuando su aspecto pudiera ser desmedrado, criaban su ternero y daban la leche mantequera que hoy es el orgullo de vacas de Centroeuropa como la Jersey y la Guernisey, que se conservan como auténticos tesoros proporcionados por la naturaleza.

Afortunadamente, nos queda la tudanca. No hay muchos ejemplares, pero estas "tasugas" pueden ser el último rescoldo que haga revivir la hoguera y el calor de lo que fue la ganadería vacuna montañesa. Este es, para mí, el motivo fundamental del librito que estamos comentando. Puede ser que la obra sirva de llamada de atención, para que los técnicos "importadores" se acuerden de

que aquí tenemos materia prima animal para ganar la partida a muchos falsos monumentos extranjeros. Sólo es necesario ponerse a trabajar, actuar con honestidad y saber de qué va.

La política de danzar por esos mundos de Dios, buscando lo mejor de otros para traérselo aquí y ganar dinero rápidamente, es propia de país pobre, subdesarrollado y sin futuro. Y que no se me diga que éstas son las características de nuestra España. Los pobres, subdesarrollados y carentes de futuro son los que buscan los caminos fáciles, sin imaginación y sin esfuerzo. Los resultados están ahí. Quienes tengan afición a la ganadería, quienes sean simplemente amantes de lo bello, lo natural y lo espontáneo, deben darse una vuelta por las deliciosas páginas de este librito que estamos comentando. Allí se van a encontrar con citas eruditas de personajes que discuten sobre el origen y ramificaciones de las razas vacunas de la montaña. Van a encontrar datos curiosos sobre las características de cada una de las estirpes. Encontrarán afirmaciones bucólicas, alusivas a los "nobles, rústicos y diminutos bovinos que constituyen la variedad de los Picos de Europa". Así, paso a paso, podrán aprender que la raza levaniega tenía un color rojizo más o menos oscuro. Con esqueletos que soportaban masas corporales de hasta 600 kilogramos. Estas páginas le dirán que la leche que producían tenía condiciones inmejorables para la elaboración de quesos y mantecas. También era exquisita su carne, ya que se trataba de razas de fácil cebo.

Los curiosos encontrarán datos interesantísimos sobre ese trozo de la nueva Castilla que fue hogar de los cántabros y que hoy conocemos también bajo la denominación genética de "la montaña". Quien lea este librito aprenderá muchas cosas sobre el territorio, la orografía, hidrografía y composición de los suelos. Obtendrá datos sobre el clima y la vegetación. Todo ello, como medio adecuado para que surgieran los animales bóvidos a que nos estamos refiriendo, que un día fueron orgullo de estas gentes laboriosas y vinculadas a sus vacas. Aunque la obra quiere hacer especial referencia a la raza tudanca, quien lea con cuidado estas páginas verá que se trata con todo cariño al resto de las razas autóctonas de Santander. Volvemos a citar aquí la levaniega, la pasiega y la campurriana.

Por mi condición de veterinario, con una vida dedicada enteramente a profundizar en el conocimiento de los animales, para aumentar sus capacidades productivas y defenderlos de las enfermedades, felicito muy efusivamente al autor de este folleto, que ha sabido recopilar un material bibliográfico disperso, para que oigan los que quieran oír y vean los que quieran ver. Llamo especialmente la atención a quienes son responsables del cuidado de nuestros ganados. Lo que la naturaleza nos dio no fue ningún capricho. Mejorándolo en lo posible debemos conservarlo, orgullosos de lo que nuestra tierra puede producir. (C. S.-G.)

AN INTRODUCTION TO HISTOTECHNOLOGY, por Geoffrey G. Brown, Ed. Appleton-Century Crofts, New York, 1978. Consideramos que en nuestra época, cuando la mayoría de los estudios histológicos y citológicos se llevan a cabo con el microscopio electrónico, interesa la edición de obras que traten de la problemática en torno a la preparación del material biológico para su observación al microscopio óptico. Este instrumento es de uso imprescindible para observaciones previas a los estudios ultraestructurales. La oportunidad de la presente resalta porque no prolifera ese tipo de bibliografía en el campo editorial y las versiones originales y las traducciones de los grandes tratados clásicos se hallan agotadas. Su aparición, un poco alejada ya en el tiempo, bien merece un somero comentario.

Tras una muy superficial descripción de la célula en el primer capítulo, se revisan uno por uno los diversos procesos a que debe someterse el material biológico para confeccionar preparaciones observables al microscopio fotónico. Da una ligera pincelada sobre la metodología seguida para los estudios ultraestructurales con el microscopio electrónico de transmisión.

La primera etapa del tratamiento del material suele ser la fijación. Y a ella dedica el primer capítulo técnico; analiza los distintos fijadores más usuales, señalando propiedades e inconvenientes. (El fijador universal no existe.) De particular interés resulta el apartado que destina a los procesos de descalcificación de las muestras. A la fijación sigue la descripción minuciosa de la inclusión en sus distintas modalidades, ilustrado con profusión de fotografías de los instrumentos automatizados que se emplean hoy en los laboratorios bien dotados.

Al capítulo de la microtomía, sin explicarnos el porqué de su inclusión, precede un estudio superficial de la parte mecánica y óptica del microscopio fotónico y de instrumentos auxiliares, como son los oculares micrométricos y las cámaras claras. Imágenes de los micrótomos convencionales y de los más modernos criostatos ilustran el apartado sobre microtomía, siendo interesante señalar la importancia, justificadísima, que da a la recomendación de la selección y cuidado de las cuchillas de acero. En esa línea pormenoriza la relación de máquinas de afilado automático existentes, instrumento imprescindible en los laboratorios histológicos bien equipados.

La relación de los defectos que pueden presentar los bloques, y por consiguiente los cortes obtenidos, tras inclusión en parafina o equivalentes, acompañada de las causas que hayan podido motivarla, así como de sus posibles enmiendas, justifican por sí mismas el libro de Brown. Tras la descripción de la técnica de la tinción y de los medios más frecuentemente empleados en el montaje de los cortes, aporta un elenco, incompleto, de los colorantes naturales y de las anilinas usados con mayor frecuencia, razonando su composición química y sus propiedades tintoriales.

Una serie de capítulos cortos cubren la explicación de las técnicas tintoriales adecuadas para los diversos tejidos, desde el epitelial al nervioso, no faltando la descripción de los métodos de Ramón y Cajal en el apartado de las impregnaciones. Diversos apéndices tratan de resolver de forma muy directa y eficaz los problemas que surgen al trabajar con técnicas histológicas, desde la limpieza del material de vidrio al calibrado correcto de los instrumentos y a la preparación de soluciones molares y normales de los reactivos de empleo más frecuente, además de la preparación de las diversas soluciones tampón. Muy oportuna nos parece la relación que figura en uno de los últimos apéndices sobre los sinónimos que tienen algunos de los colorantes más usuales, para finalizar con una lista de las casas comerciales suministradoras de los instrumentos y reactivos citados en el texto.

Aunque la obra no aporte nada espectacularmente nuevo, está bien documentada. Al final de cada capítulo incorpora una relación bibliográfica sobre el tema altamente cualificada, pero sin ánimo de mostrarse exhaustiva. La obra se ordena de acuerdo con un esquema válido para quien se inicia en la técnica histológica; estaría fuera de lugar pretender compararla con las guías formularios clásicos de Romeis o Langeron, ni con las más recientes de Gabe, Martoja o Nezeloff, que el investigador profesional conoce bien. Vale la pena destacar, por último, el valioso inventario fotográfico de los modernos instrumentos utilizados en inclusiones y los microtomos. (M. D.)

Bibliografía

Los lectores interesados en una mayor profundización de los temas expuestos pueden consultar los trabajos siguientes:

REPARACION DEL MATERIAL GENETICO

MOLECULAR MECHANISMS FOR REPAIR OF DNA. Dirigido por Philip C. Hanawalt y Richard B. Setlow. Plenum Press, 1975.

ULTRAVIOLET MUTAGENESIS AND INDUCIBLE DNA REPAIR IN *ESCHERICHIA COLI*. Evelyn M. Witkin, en *Bacteriological Reviews*, vol. 40, págs. 869-907, 1976.

DNA: REPLICATION AND RECOMBINATION. Cold Spring Harbor Symposia on Quantitative Biology, vol. 43, 1978.

CAUSAS DE LA DIABETES

RECEPTORS FOR PEPTIDE HORMONES: NEW INSIGHTS INTO THE PATHOPSYCHOLOGY OF DISEASE STATES IN MAN. Richard C. Eastman y Jeffrey S. Flier en *Annals of Internal Medicine*, vol. 86, nº. 2, págs. 205-219; febrero, 1977.

TYPE I DIABETES MELLITUS. A. G. Cudworth in *Diabetologia*, vol. 14, n.º 5, págs. 281-291; mayo, 1978.

VIRUS-INDUCED DIABETES MELLITUS: ISOLATION OF A VIRUS FROM THE PANCREAS OF A CHILD WITH DIABETIC KETOACIDOSIS. Ji-Won Yoon, Marshall Austin, Takashi Onodera y Abner Louis Notkins en *The New England Journal of Medicine*, vol. 300, n.º 21, págs. 1173-1179; 24 de mayo de 1979.

ALEACIONES CON MEMORIA DE LA FORMA

NITINOL HEAT ENGINES. Ridgway Banks en *Shape Memory Effects in Alloys*, dirigido por Jeff Perkins. Plenum Press, 1975.

SHAPE MEMORY EFFECTS AND APPLICATIONS: AN OVERVIEW. Walter S. Owen en *Shape Memory Effects in Alloys*, dirigido por Jeff Perkins. Plenum Press, 1975.

TEORIA NEUTRALISTA DE LA EVOLUCION MOLECULAR

THEORETICAL ASPECTS OF POPULATION GENETICS. Motoo Kimura y Tomoka Ohta. Princeton University Press, 1971.

MOLECULAR POPULATION GENETICS AND EVOLUTION. Masatoshi Nei. North-Holland Publishing Co., 1975.

STATISTICAL STUDIES ON PROTEIN POLYMORPHISM IN NATURAL POPULATIONS, I: DISTRIBUTION OF SINGLE LOCUS HETEROZYGOSITY. Paul A. Fuerst, Ranajit Chakraborty y Masatoshi Nei en *Genetics*, vol. 86, n.º 2, parte 1, págs. 455-483; junio, 1977.

LACK OF EXPERIMENTAL EVIDENCE FOR FREQUENCY-DEPENDENT SELECTION AT THE ALCOHOL DEHYDROGENASE LOCUS IN *DROSOPHYLA MELANOGASTER*. Hiroshi Yoshimaru y Terumi Mukai en *The Proceedings of the National Academy of Sciences of the United States of America*, vol. 76, n.º 2, págs. 876-878; febrero, 1979.

MODEL OF EFFECTIVELY NEUTRAL MUTATIONS IN WHICH SELECTIVE CONSTRAINT IS INCORPORATED. Motoo Kimura en *The Proceedings of the National Academy of Sciences of the United States of America*, vol. 76, n.º 7, págs. 3440-3444; julio, 1979.

LAS GALAXIAS PRIMITIVAS

ARE YOUNG GALAXIES VISIBLE? R. B. Partridge y P. J. E. Peebles en *Astrophysical Journal*, vol. 147, n.º 3, págs. 868-886; marzo, 1967.

THE OPTICAL APPEARANCE OF MODEL PRIMEVAL GALAXIES. David L. Meier en *Astrophysical Journal*, vol. 207, n.º 2, págs. 343-350; 15 de julio de 1976.

YOUNG GALAXIES, QUASARS AND THE COSMOLOGICAL EVOLUTION OF EXTRAGALACTIC RADIO SOURCES. M. S. Longair y R. A. Sunyaev en *Radio Astronomy and Cosmology: International Astronomical Union Symposium No. 74*, dirigido por David L. Jauncey. D. Reidel Publishing Company, 1977.

RECENT THEORIES OF GALAXY FORMATION. J. Richard Gott III en *Annual Review of Astronomy and Astrophysics*, vol. 15, págs. 235-266; 1977.

ECOLOGIA DEL ESCARABAJO ESTERCOLERO AFRICANO

WHY HAVE SOME ANIMALS EVOLVED TO REGULATE A HIGH BODY TEMPERATURE? Bernd Heinrich en *American Naturalist*, vol. 111, n.º 980, págs. 623-640; julio-agosto, 1977.

ENDOTHERMY IN AFRICAN DUNG BEETLES DURING FLIGHT, BALL MAKING, AND BALL ROLLING. George A. Bartholomew y Bernd Heinrich en *The Journal of Experimental Biology*, vol. 73, n.º 2, págs. 65-83; abril, 1978.

TEORIA CUANTICA Y REALIDAD

CAN QUANTUM-MECHANICAL DESCRIPTION OF PHYSICAL REALITY BE CONSIDERED COMPLETE? A. Einstein, B. Podolsky y N. Rosen en *Physical Review*, vol. 47, n.º 10, págs. 777-780; 15 de mayo de 1935.

ON THE EINSTEIN PODOLSKY ROSEN PARADOX. J. S. Bell en *Physics*, vol. 1, n.º 3, págs. 195-200; noviembre/diciembre, 1964.

FOUNDATIONS OF QUANTUM MECHANICS. Dirigido por B. d'Espagnat. Academic Press, 1971.

EXPERIMENTAL CONSEQUENCES OF OBJECTIVE LOCAL THEORIES. John F. Clauser y Michael A. Horne en *Physical Review D*, vol. 10, n.º 2, págs. 526-535; 15 de julio de 1974.

USE OF INEQUALITIES FOR THE EXPERIMENTAL TEST OF A GENERAL CONCEPTION OF THE FOUNDATIONS OF MICROPHYSICS. B. d'Espagnat en *Physical Review D*, vol. 11, n.º 6, págs. 1424-1435, 15 de marzo de 1975; parte 2, vol. 18, n.º 2, págs. 349-358, 15 de julio de 1978.

CONCEPTUAL FOUNDATIONS OF QUANTUM MECHANICS. Bernard d'Espagnat. W. A. Benjamin, Inc., 1976.

BELL'S THEOREM: EXPERIMENTAL TESTS AND IMPLICATIONS. John F. Clauser y Abner Shimony en *Reports on Progress in Physics*, vol. 41, n.º 12, págs. 1881-1927; diciembre, 1978.

UN ESTABLECIMIENTO NEOLITICO Y DE LA EDAD DE HIERRO EN UNA COLINA INGLESA

IRON AGE COMMUNITIES IN BRITAIN. Barry Cunliffe. Routledge & Kegan Paul, 1974.

THE IRON AGE IN LOWLAND BRITAIN. D. W. Harding. Routledge & Kegan Paul, 1974.

HILLFORTS: LATER PREHISTORIC EARTHWORKS IN BRITAIN AND IRELAND. Dirigido por D. W. Harding. Academic Press, 1976.

JUEGOS MATEMATICOS

ALGORITHMIC INFORMATION THEORY. G. J. Chaitin en *IBM Journal of Research and Development*, Vol. 21, n.º 4, págs. 350-359, julio, 1977, y errata, vol. 21, n.º 5, pág. 496, septiembre, 1977.

TALLER Y LABORATORIO

THE OPTICS OF FLAMES. Felix J. Weinberg. Butterworth & Co., 1963.

FLAME STRUCTURE. R. M. Fristrom y A. A. Westenberg. McGraw-Hill Book Company, 1965.

FLAMES: THEIR STRUCTURE, RADIATION AND TEMPERATURE. A. G. Gaydon y H. G. Wolfhard. Chapman and Hall, Ltd., 1970.